On factorisation forests

And some applications

Thomas Colcombet

Cnrs/Irisa
thomas.colcombet@irisa.fr

Keywords: Formal languages, semigroups, infinite words, trees, monadic second-order logic, infinite structures.

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. The purpose of this paper is to illustrate the importance of this approach in the context of automata over infinite words and trees.

We extend the theorem of factorisation forest in two directions: we show that it is still valid for any word indexed by a linear ordering; and we show that it admits a deterministic variant for words indexed by well-orderings. A byproduct of this work is also an improvement on the known bounds for the original result.

We apply the first variant for giving a simplified proof of the closure under complementation of rational sets of words indexed by countable scattered linear orderings. We apply the second variant in the analysis of monadic second-order logic over trees, yielding new results on monadic interpretations over trees. Consequences of it are new caracterisations of prefix-recognizable structures and of the Caucal hierarchy.

1 Introduction

Factorisation forests were introduced by Simon [24]. 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 exponent 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 [24]. This theorem also allows a constructive proof of Brown's lemma on locally finite semigroups [6]. It is also used in the caracterisation of subfamilies of the regular languages, for instance the polynomial closure of varieties in [17]. Or to give general caracterisations of finite semigroups [20]. In this last paper, the result is applied for proving McNaughton's determinisation results of automata over infinite words [15]. In the context of languages of infinite words indexed by ω, it has also been used in a complemetation procedure [5] extending Buchi's lemma [1].

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 all of its proofs rely on the use of Green's relations: Green's relations form an extremely 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 mentionned above and also in this paper, this theorem as already important applications to automata theory. This is why this theorem is worth being advertised outside the semigroup community as a major tool in automata theory.

The technical contribution of the paper is an investigation of the potential use of factorisation forests in broader contexts than finite words. An important objective is to be able to apply this theorem on infinite words, and on trees instead of words. Those attempts are incarnated by two new variants of the theorem. As a byproduct we improve the known bounds of the original result (in particular on the previous improvement [13]).

We also provide some applications of those results. We give a new proof of the result of Carton and Rispal showing the closure under comlementation of rational languages of words with countable scattered linear domain [10]. We use the other variant of the theorem for proving a decomposition result for monadic interpretations (in fact the application of a technique that we call compaction). This yields new caracterisations of prefix recognisable structures and of the Caucal hierarchy.

However, the applications of those results go beyond the one proposed here. In paricular, let us mention the work of Blumensath [4] who applies the deterministic variant of the theorem presented here for giving a new proof of Rabin's theorem [21]. The theorem of Rabin states that the monadic theory of the infinite binary tree is decidable. Different proofs have been proposed for this result so far, all relying on the use of automata theory, and most of them on the use of parity games (see [25] for a survey). For the simpler theory of the naturals with successor — originally proved by Buchi [1] — another proof technique is known: the compositional method of Shelah [23]. In this seminal paper, Shelah asks whether there exists a proof of Rabin's theorem along the same lines. Blumensath [4] answers to this longstanding open question positively.

The content of the paper is organised as follows. Section 2 is dedicated to definitions. Section 3 present the original theorem of factorisation forests as well as two less standard presentations of it. We also introduce in this section the notion of a ramseyan split, which is central in the remainder of the paper. In Section 4 we provide the first extension of the theorem, the extension to all complete linear orderings. Section 5 is dedicated to the application of this extension to the complementation of automata over countable scattered linear orderings. In Section 6 we provide the second extension of the theorem, to ordinals only this time, but with an extra property of determinism. Finally, in Section 7, we develop the technique of compaction and use it for providing a new decomposition result for monadic interpretations applied to trees. We also show how this impacts on the theory of infinite structures.

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 $y < x$. 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 $\beta \subseteq \alpha$. We omit the ordering relation $<$ below unless necessary, and just say that L is a linear ordering. An *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. 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 to $<_1$ on L_1, to $<_2$ 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 *well ordered* if every nonempty subset has a minimal element. It 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 *cut* in 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 over L and by $\overline{\alpha}^*$ the set $\overline{\alpha} \setminus \{\bot, \top\}$. An important remark is that $\overline{\alpha}$ and $\overline{\alpha}^*$ are complete linear orderings.

Cuts can be thought as new elements located between the elements of L: given $x \in L$, $x^- = (]-\infty, x[, [x, +\infty[)$ 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 L (resp. not the maximal element and not of the form x^-).

Two linear orderings $\alpha = (L, <)$ and $\beta = (L', <')$ are *isomorphic* if there exists a bijection f from L onto L' such that for every x, y in L, $x < y$ iff $f(x) <' f(y)$. In this case, we also say that $(L, <)$ and $(L', <')$ have *the same order type.* This is an equivalence relation on the class of linear orderings. We denote by ω, ω^*, ζ the order types of respectively $(\mathbb{N}, <)$ (the naturals), $(-\mathbb{N}, <)$ (the nonpositive integers) and $(\mathbb{Z}, <)$ (the integers). The order type of a well-ordering is called an *ordinal.* Below, we do often not distinguish between a linear ordering and its type. This is safe since all the construction we perform are isomorphism invariant.

The interested reader can find in [22] additional material on linear orderings.

2.2 Words

We use a generalized version of words: words indexed by a linear ordering. 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. 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. This construction is naturally generalized 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.

2.3 Semigroups and additive labellings

For a thorough introduction to semigroups, we refer the reader to [14,18,19]. 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 $\varphi_u(x,y)$ is $\varphi(u|_{]x,y[})$. I.e. $\varphi_u(x,y)$ is the image by φ of the factor of u located between x and y. We denote by φ_u^* the mapping φ_u restricted to $\overline{\alpha}^*$.

2.4 Structures, graphs, trees, logics

Relational structures Let us first remark that the definitions presented here are useless before Section 6, have marginal consequences in Section 6, and are of real interest only for Section 7.

A *relational structure* $(\mathcal{U},R_1,\dots,R_n)$ is a set $\mathcal{U}$, called the *universe*, together with *relations* $R_1,\dots,R_n$ of fixed finite arity over $\mathcal{U}$. Each relation R has a *name* that we write R itself. The *signature* of a structure contains the names involved together their arity. A *graph* is a relational structure for which the relations have arity 1 and one relation of arity 2. The elements of the universe are called *vertices*, the unary relations are called *label relations*, and the binary relations is called the *edge relation*. A *path* is a finite sequence of vertices such that two successive vertices are in relation by the edge relation. The first vertex is called the *origin* of the path, and the last vertex the *destination*.

Linear orderings can be naturally represented as graphs: $(L,<)$ can be seen as a graph of vertices L, with an edge between x and y iff $x<y$. For a linear ordering $\alpha=(L,<)$ and a finite alphabet $A=\{a_1,\dots,a_n\}$, an α-word u is the graph $(L,<,a_1,\dots,a_n)$ obtained from the graph of the linear ordering by setting a_i to be interpreted as $u^{-1}(a_i)$; the set of positions in the word corresponding to letter a_i.

A *tree* t is a graph such that there is only one edge relation, called the *ancestor relation* and denoted $\sqsubseteq$, satisfying:

- the relation $\sqsubseteq$ is an order,
- there is a minimal element for $\sqsubseteq$, called the *root*,
- for every u, the set $\{v \;:\; v\sqsubseteq u\}$ is an ordinal of length at most ω.

The vertices of a tree are called *nodes*. Maximal chains of nodes in a tree are called *branches*.

Warning: The trees are *not* defined by a 'direct successor' relation, but rather by the ancestor relation. This has major impact on the logical side: all the logics we use below can refer to the ancestor relation, and it is well-known that first-order logic using this ancestor relation is significantly more expressive over trees than first-order logic with access to the successor of a node only. The results would fail if the ancestor relation was not used.

A particular tree will play a special role below. The *complete binary tree* has as universe $\{0,1\}^*$, as ancestor relation the prefix relation, and has two unary relations, $0=\{0,1\}^*0$ and $1=\{0,1\}^*1$. We call the relation 0 the *left-child relation*, while 1 is the *right-child relation*. We denote by Δ_2 the complete binary tree.

One constructs a tree from a graph by unfolding. Given a graph G and one of its vertices v, the *unfolding* of G from v is the tree which has as nodes the all paths with origin v, as ancestor relation the prefix relation over paths, and such that a path π is labelled by a in the unfolding iff its destination is labelled by a in the graph.

Logics For defining first-order logic, we need to have at our disposal a countable set of *first-order variables* $x,y,\dots$ to pick from. The *atomic formulæ* are $R(x_1,\dots,x_n)$ for $x_1,\dots,x_n$ first-order variables and R the name of relation of arity n; given two first-order variables x, y, $x = y$ is also an atomic formula. *First-order* logic formulæ are made out of these atomic formulæ, combined with the boolean connectives $\vee, \wedge, \neg$, and the first-order quantifiers $\exists x$ and $\forall x$. For monadic logic, we need furthermore a countable set of *monadic variables* $X, Y, \ldots$ *Monadic (second-order) formulæ* are defined as first-order formulæ, but further allow the use of monadic quantifiers $\exists X$, $\forall X$, and of a membership atomic formula $x \in X$, where x is a first-order variable and X a monadic one. For first-order as well as monadic formulæ we use the standard notion of *free variables*. A formula without free variables is called a *closed* formula.

We denote by $\mathcal{S} \models \phi$ the fact, for a closed formula ϕ and a structure $\mathcal{S}$, that the formula is true over the structure $\mathcal{S}$. The formal definition uses the standard semantic, the value of first-order variables ranging over elements of the universe of the structure, while monadic variables take as values subsets of the universe. We say that $\mathcal{S}$ is a *model* of ϕ, or that ϕ is *satisfied* over $\mathcal{S}$. When the structure is obvious from the context, we simply state that ϕ is satisfied. We also allow ourselves to use formulæ like $\phi(x_1, \ldots, x_n)$ to denote that the free-variables of ϕ are among $\{x_1, \ldots, x_n\}$. Then given elements $u_1, \ldots, u_n$ in the universe of a structure $\mathcal{S}$, we write $\mathcal{S} \models \phi(u_1, \ldots, u_n)$ if the formula ϕ is true over the structure $\mathcal{S}$, using the valuation which to each x_i associates u_i.

A relational structure $\mathcal{S}$ has a *decidable L-theory* (where L is either first-order or monadic), if there is an algorithm which, given a formula ϕ of the logic L, answers whether $\mathcal{S} \models \phi$ or not.

Interpretations An interpretation is an operation defined by logic formulæ that defines a structure inside another one. An *interpretation* is given as a tuple

$$\mathcal{I} = (\delta(x), \phi_1(x_1, \ldots, x_{|R_1|}), \ldots, \phi_k(x_1, \ldots, x_{|R_K|}))$$

where $\delta(x), \phi_1(x_1, \ldots, x_{|R_1|}), \ldots, \phi_k(x_1, \ldots, x_{|R_K|})$ are formulæ of corresponding free variables. The interpretation is *first-order* if the formulæ are first-order and *monadic* if the formulæ are monadic.

Given a structure $\mathcal{S}$ of universe $\mathcal{U}$, $\mathcal{I}(\mathcal{S})$ is the structure of universe

$$\mathcal{U}_{\mathcal{I}(\mathcal{S})} = \{u \in \mathcal{U} \ : \ \mathcal{S} \models \delta(u)\} \ ,$$

and such that the interpretation of R_i is

$$\{(u_1, \ldots, u_{|R_i|}) \in \mathcal{U}_{\mathcal{I}(\mathcal{S})}^{|R_i|} \ : \ \mathcal{S} \models \phi_i(u_1, \ldots, u_{|R_i|})\}.$$

A special case of interpretation is the *marking*. A marking replicates the structure, and adds some new unary relations on it.

3 Factorisation forest theorem: various presentations for the standard case

In this section, we present the theorem of factorisation forest. We first give the original statement in Section 3.1. The in Section 3.2, we provide another equivalent presentation in terms of regular expressions; possibly the most natural one. In Section 3.3, we introduce the notion of a split, and use it for a third formalisation of the result. This notion is the one used in the extensions of the factorisation forest theorem we provide below.

3.1 Factorisation forest theorem

Fix an alphabet A and a semigroup morphism φ from A^+ to a finite semigroup $(S, .)$. A *factorisation tree* of a word $u \in A^+$ is an ordered unranked tree in which each node is either a leaf

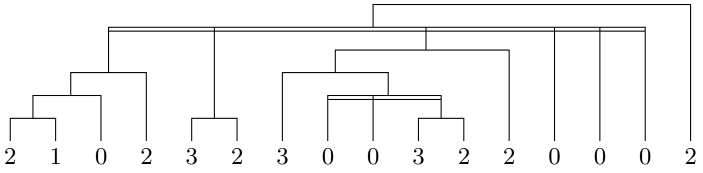

Fig. 1. A factorisation tree

labeled by a letter, or an internal node, and such that the word obtained by reading the leaves from left to right (the yield) is u. The *height* of the tree is defined as usual, with the convention that the height of a tree restricted to 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 $u = 210232300322002$ (u is the yield of the tree). In this drawing, internal nodes appear as horizontal lines. Double line correspond to case 3 in the description of ramseyanity.

The theorem of factorisation forests is then the following.

Theorem 1 (factorisation forests). *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 [24], with a bound of $9|S|$. An improved bound of $7|S|$ is provided by Chalopin and Leung [13]. The value of $3|S|$ is a byproduct of the present work.

3.2 A variant via regular expressions

The use of factorisation trees gave the name of factorisation forests to the theorem. But it is sometime very convenient to use another formalisation in terms of regular expressions. This presentation is new (to the knowledge of the author), but its simplicity makes it worth to be advertised. Let A be an alphabet, φ a semigroup morphism from A^+ to some semigroup S, and E be a regular expression over the alphabet A. E is *φ-ramseyan* if for each occurence L^* of the Kleene star in E, L is mapped to $\{e\}$ by φ, for e an idempotent in S.

Example 2. Let S be $\mathbb{Z}/2\mathbb{Z}$ with the addition, A be $\{0, 1\}$ and φ be the morphism from A^+ to S sending each letter to its value modulo 2. The expression $0(0 + 10^*1)^* + 10^*1(0 + 10^*1)^*$ is φ-ramseyan and evaluates to $\varphi^{-1}(0)$.

Theorem 2 (variant of factorisation forests). *For every alphabet A, finite semigroup $(S, .)$, semigroup morphism φ from A^+ to S and x in S, there exists a φ-ramseyan regular expression E_x evaluating to $\varphi^{-1}(x)$.*

Proof. By induction on k, for every x in S, let the φ-ramseyan regular expression E_x^k be:

$$E_x^0 = \varphi^{-1}(x) \cap A \ , \qquad\qquad E_x^{k+1} = E_x^k + \sum_{yz=x} E_y^k E_z^k + \sum_{e^2=e=x} (E_e^k)^* \ .$$

On can show by induction on k that for all $x \in S$, E_x^k evaluates to the set of words in $\varphi^{-1}(x)$ possessing a factorisation tree of height k. This proof, for both directions of the inclusion, is a direct application of the definitions. Then, by Theorem 1, $E_x^{3|S|}$ evaluates to $\varphi^{-1}(x)$. □

The interest of Theorem 2 is that it allows to perform proofs by induction on the structure of ramseyan regular expressions. By the following refinement, we can derive complexities when using this technique.

Property 1 (refinement of Theorem 2). The height of the regular expression E_x is at most $3|S|+1$, counting 0 for the operator $+$, and 1 for the concatenation, the Kleene star and constants. The regular expression E_x contains at most $6|S|^2$ distinct subexpressions, at most $3|S|^2$ distinct subexpressions without the $+$-operator at the root.

Those bounds are obtained from the last variant, Theorem 3.

3.3 A variant via ramseyan splits

The third equivalent presentation to the theorem of factorisation forests uses the notion of ramseyan splits. One way to see a split is as a form of presentation of a tree. This formalisation naturally extends to infinite words, and is very natural to use in automata theoretic constructions. The extensions of the theorem proposed in the remaining of the paper use this definition.

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 of the semigroup.

Example 3. 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 $s(x)$:

$$\begin{array}{c}
1\quad 3\quad 2\quad 2\quad 1\quad 2\quad 1\quad 2\quad 2\quad 2\quad 3\quad 2\quad 1\quad 1\quad 1\quad 1\quad 2\\
|\,2\,|\,1\,|\,0\,|\,2\,|\,3\,|\,2\,|\,3\,|\,0\,|\,0\,|\,3\,|\,2\,|\,2\,|\,0\,|\,0\,|\,0\,|\,2\,|
\end{array}$$

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 3. *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 of height at most $|S|$.*

The proof of this result is postponed to Section 4.2, as the proof is a simplification of the proof of its extension Theorem 4.

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 $u\in A^+$. 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 φ^*_u,
- every ramseyan split of height k of φ^*_u can be turned into a factorisation tree of height at most $3k$ of u.

Using this last argument and Theorem 3, we directly obtain a proof of Theorem 1 with the announced bound of $3|S|$. Using similar arguments, one obtains the bounds of Property 1.

4 Extension of the factorisation forest theorem to infinite words

The contribution of this section is an extension of Theorem 3 to complete linear orderings.

Theorem 4. *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 3, 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 an ordinal, proves Theorem 3.

The remaining of the section is devoted to the proof of Theorem 4, as well as its ordinal version, Theorem 3. We start in Section 4.1 by establishing some elementary topological lemmas relative to complete linear orderings. Then, in Section 4.2, we give successively a proof of both Theorems 3 and 4.

4.1 On linear orderings

The subject of this section is to provide preparatory lemmas on linear orderings. Namely Lemmas 1 and 3. This Section is not relevant for the simpler proof of Theorem 3.

We consider here a binary relation R over a linear ordering α. The statement $R(x,y)$ can be thought as meaningful only for $x<y$, in the sense that we do not take into account the value of R elsewhere. We say that a binary relation R over α is *upward closed* if for every $x\leq x'<y'\leq y$, $R(x',y')$ implies $R(x,y)$.

Lemma 1. *Let α be a complete linear ordering, and R be an upward closed relation over α. There exists $\gamma\subseteq\alpha$ such that for every $x<y$ in α,*

- *if $R(x,y)$ then $[x,y]\cap\gamma$ is nonempty,*
- *if $]x,y[\cap\gamma$ contains two distinct elements, then $R(x,y)$.*

Let us first remark that if Lemma 1 holds for some linear ordering α, then it is also true for every convex subset of α. For this reason, we can safely add a new minimal element $\bot'$ and maximal element $\top'$ to α, such that for every x in α, $R(\bot',x)$ and $R(x,\top')$. Define now for $x\in\alpha$,

$$\begin{aligned} l(x) &= \sup\{y \,:\, \forall z>x.\ R(y,z)\}\ , \\ \text{and}\quad r(x) &= \inf\{z \,:\, \forall y<x.\ R(y,z)\}\ . \end{aligned}$$

Thanks to the adjunction of $\bot'$ and $\top'$, l and r are defined everywhere but for the minimal and maximal elements respectively.

Fact 2. *The following holds.*

1. *Both l and r are nondecreasing.*

2. *For every x, $l(x) \leq x \leq r(x)$.*
3. *For every x, $l(x) = x$ iff $r(x) = x$.*
4. *For every x, $r^\omega(x) = \sup\{r^n(x) : n \in \mathbb{N}\}$ and $l^\omega(x) = \inf\{l^n(x) : n \in \mathbb{N}\}$ are fixpoints of both l and r.*
5. *For every x, y, z, if $x < z \leq r(z) < y$ then $R(x, y)$.*
6. *For every x, y, z, if $z < x < y < r(z)$ then $\neg R(x, y)$.*

Proof. Items 1,2,5 and 6 follow from the definition.

For item 3. By upward closure of R, $l(x) = x$ iff for every $y < x$ and $z > x$, $R(y, z)$, iff $r(x) = x$.

For item 4. Let $y = r^\omega(x)$. By item 2, we have $y \leq r(y)$. We have to prove $r(y) \leq y$. Let x_n be $r^n(x)$. If $x_{n+1} = x_n$ for some n, then $y = x_n = r(x_n) = r(y)$. Else $x_0 < x_1 < \cdots < y$. It follows by definition of r that for all n, $R(x_n, y)$. This implies $r(y) = y$. □

We can now prove Lemma 1.

Proof. Set Fix to be the set of fixpoints of r (equivalently, l). Define the equivalence relation $\sim$ by $x \sim y$ if $x = y \in \text{Fix}$ or $[x, y] \cap \text{Fix}$ is empty. This relation induces two kind of equivalence classes: singletons consisting of a single fixpoint, or maximal intervals containing no fixpoint.

Let C be an equivalence class of $\sim$. If $C = \{x\}$ for $x \in \text{Fix}$, set $\gamma(C)$ to be C. Else, C is an interval. Fix an element x_C in C, set x_C^n to be $r^n(x_C)$ for $n \geq 0$ and x_C^{-n} be $l^n(x)$ for $n \geq 0$ (both definitions coincide for $n = 0$ with $x_C^0 = x_C$). By induction and using fact 2, one easily shows that for every n, both x_C^n and x_C^{-n} belong to $\text{Fix} \cup C$. Let $\gamma(C)$ be $\{x_C^n : n \in \mathbb{Z}, x_C^n \notin \text{Fix}\}$. According to the previous remark $\gamma(C) \subseteq C$.

We now define γ to be the union of $\gamma(C)$ for C ranging over equivalence classes of $\sim$. Let us prove that this γ satisfies the conclusion of the lemma.

Let $x < y$ be in α such that $]x, y[\cap \gamma$ contains two distinct elements. If $]x, y[$ contains two elements $x' < y'$ nonequivalent for $\sim$, there is a fixpoint in $[x', y'] \subseteq]x, y[$. It follows by Fact 2 that $R(x, y)$. Else $]x, y[$ is included in some equivalence class C of $\sim$. Thus, the two elements in $]x, y[$ are of the form x_C^n and x_C^m for $n < m$. Since $x_C^n < x_C^{n+1} \leq x_C^m$, $x_C^{n+1} = r(x_C^n)$ belongs to $]x, y[$. By Fact 2, $R(x, y)$.

Let $x < y$ be in L such that $R(x, y)$. If $x \not\sim y$ then by definition $\text{Fix} \cap [x, y]$ is nonempty. And since $\text{Fix} \subseteq \gamma$, $[x, y] \cap \gamma$ is nonempty. Else $x \sim y$. Let C be the equivalence class containing both x and y. If $x_C \in [x, y]$, then x_C witnesses the nonemptyness of $\gamma \cap [x, y]$. Else either $x > x_C$ or $y < x_C$. The two cases are symmetric. Let us treat the case $x > x_C$. By Fact 2, $r^\omega(x_C) \in \text{Fix}$, and as $x_C \sim x$, $x < r^\omega(x_C)$. Hence, there exists some n in $\mathbb{N}$ such that $x_C^n = r^n(X_C) \geq x$. Let n be the least such natural. We have $x_C^{n-1} < x$, and by monotonicity (Fact 2) $x_C^n \leq r(x)$. Overall $x_C^n \in [x, r(x)]$. Furthermore by Fact 2, $r(x) \leq y$. This witnesses $x_C^n \in \gamma \cap [x, y]$. □

We will also require the following lemma[1].

Lemma 3. *For every linear ordering α and every natural k, there exists a mapping $c : \alpha \to \{0, \ldots, k-1\}$ such that for every $x < y$ in α with $c(x) = c(y)$, $c([x, y]) = \{0, \ldots, k-1\}$.*

Proof. Let $[k]$ denote $\{0, \ldots, k-1\}$ We first show the result for a dense linear ordering β. Consider the set M of partial mappings c from β to $[k]$ such that for every $x < y$ with $c(x) = c(y)$ defined, either c is injective when restricted to $[x, y[$, or $c([x, y]) = [k]$. Those mappings are ordered by $c \subseteq c'$ if the domain of c' contains the domain of c, and c coincides with c' over its domain. Consider now a chain $(c_i)_{i \in I}$ of elements in M. It has an upper bound b defined

[1] In fact, the weaker result needed is the existence of a mapping $c : \alpha \to \{0, \ldots, k-1\}$ such that for all $x < y$ in α with $c(x) = c(y) = 0$, $c([x, y]) = \{0, \ldots, k-1\}$. It happens to be much easier to establish than Lemma 3.

by $b(x) = c_i(x)$ if there is some i such that $c_i(x)$ is defined, else $b(x)$ is undefined. It is easy to check that b belongs also to M. By Zorn's lemma, there exists a maximal element m in M. Assume m is not defined in say, x. Let Y be the set of elements y such that m is not defined over $[\min(x,y), max(x,y)]$. By definition, $x \in Y$. There are four cases depending on whether x is the minimal (resp. the maximal) element of Y. If x is neither the minimal nor the maximal element, this means there exists $y < x < z$ in Y. By density, we can construct a ζ-indexed growing sequence $(x_i)_{i \in \mathbb{Z}}$ included in Y. Define then m' to coincide everywhere with m, but over the x_i's, where $m'(x_i)$ is set to be the remainder of i modulo k. By construction m' belongs to M, contradicting the maximality of m. If Y is $[x]$, set m' to coincide everywhere with m but for x, where $m'(x) = 0$. Once more, m' belongs to M, this time by remarking that every value in $[k]$ is mapped by m infinitely close to the left and to the right of x. This contradicts the maximality of m. The other possibilities for Y are just combinations of the two above. Hence m has to be defined everywhere, which means by density of β that the conclusion of the lemma holds for every dense linear ordering.

At this point, the easiest way to conclude the proof is to prove for every n in $[k]$ and every scattered nonempty linear ordering β, that there exists a mapping $c_{\beta,n}$ satisfying the conclusion of the lemma, such that $c_{\beta,n}^{-1}(n)$ is nonempty. This can be easily done with the help of Hausdorff's theorem (see e.g. chapter 5 in [22]). Then, one uses the fact that every linear ordering α is a dense sum of scattered linear orderings (Theorem 4.9 in [22]), i.e:

$$\alpha = \sum_{x \in \gamma} \beta_x \quad \text{with } \gamma \text{ dense, and all the } \beta_x \text{ are scattered and pairwise disjoint.}$$

Then, using the case of a dense linear ordering above, we have a mapping d from γ to $[k]$ satisfying the conclusion of the lemma. Define now c over α by $c(x) = c_{\beta,d(\beta)}(x)$ for $\beta \in \gamma$ with $x \in \beta$. This mapping c fulfills the conclusion of the lemma. □

4.2 Proof of the statement

We assume here the reader used to standard semigroup theory, and in particular Green's relations. The reader can refer to [14,18,19] for a presentation of the subject. Some definitions and facts are presented below.

Below, σ denotes the additive labelling from the complete linear ordering α to the finite semigroup $(S, .)$ of Theorem 4. We denote by β a subordering of α. We slightly abuse the notation, and write (β, σ) for $(\beta, \sigma|_\beta)$ in which $\sigma|_\beta$ is the additive labelling obtained by restricting σ to β. We also denote by $\sigma(\beta)$ the set $\{\sigma(x,y) \ : \ x < y, \ x, y \in \beta\}$.

Facts about finite semigroups and Green's relations

We recall some definitions here, and gather some standard facts concerning finite semigroups.

Given a semigroup S, S^1 denotes the monoid S itself if S is a monoid, or the monoid S augmented with a new neutral element 1 otherwise, thus making S a monoid.

The Green's relation are defined by:

$$\begin{array}{llllll}
a \leq_{\mathcal{L}} b & \text{if} & a = cb \text{ for some } c \text{ in } S^1 & \quad a \mathrel{\mathcal{L}} b & \text{if} & a \leq_{\mathcal{L}} b \text{ and } b \leq_{\mathcal{L}} a \\
a \leq_{\mathcal{R}} b & \text{if} & a = bc \text{ for some } c \text{ in } S^1 & \quad a \mathrel{\mathcal{R}} b & \text{if} & a \leq_{\mathcal{R}} b \text{ and } b \leq_{\mathcal{R}} a \\
a \leq_{\mathcal{J}} b & \text{if} & a = cbc' \text{ for some } c, c' \text{ in } S^1 & \quad a \mathrel{\mathcal{J}} b & \text{if} & a \leq_{\mathcal{J}} y \text{ and } b \leq_{\mathcal{J}} a \\
a \leq_{\mathcal{H}} b & \text{if} & a \leq_{\mathcal{L}} b \text{ and } a \leq_{\mathcal{R}} b & \quad a \mathrel{\mathcal{H}} b & \text{if} & a \mathrel{\mathcal{L}} b \text{ and } a \mathrel{\mathcal{R}} b
\end{array}$$

Fact 4. *Let a, b, c be in S. If $a \mathrel{\mathcal{L}} b$ then $ac \mathrel{\mathcal{L}} bc$. If $a \mathrel{\mathcal{R}} b$ then $ca \mathrel{\mathcal{R}} cb$. For every a, b in S, $a \mathrel{\mathcal{L}} c \mathrel{\mathcal{R}} b$ for some c iff $a \mathrel{\mathcal{R}} c' \mathrel{\mathcal{L}} b$ for some c'.*

As a consequence of the last equivalence, one defines the last of Green's relations:

$$\begin{aligned} a\,\mathcal{D}\,b \quad & \text{if} \quad a\,\mathcal{L}\,c\,\mathcal{R}\,b \text{ for some } c \text{ in } S \\ & \text{if} \quad a\,\mathcal{R}\,c'\,\mathcal{L}\,b \text{ for some } c' \text{ in } S \end{aligned}$$

The key result being (here the hypothesis of finiteness of S is mandatory):

Fact 5. *$\mathcal{D}$=$\mathcal{J}$.*

For this reason, we refer from now on only to $\mathcal{D}$ and not $\mathcal{J}$. However, we will use the preorder $\leq_{\mathcal{J}}$ (which is an order over the $\mathcal{D}$-classes).

An elemement a in S is called *regular* if $asa = a$ for some s in S. A $\mathcal{D}$-class is *regular* if all its elements are regular.

Fact 6. *A $\mathcal{D}$-class D is regular, iff it contains an idempotent, iff every $\mathcal{L}$-class in D contains an idempotent, iff every $\mathcal{R}$-class in D contains an idempotent, iff there exists a, b in D such that $ab \in D$.*

Fact 7. *For every a, b in D such that $ab \in D$, $a\,\mathcal{R}\,ab$ and $b\,\mathcal{L}\,ab$. Furthermore, there is an idempotent e in D such that $a\,\mathcal{L}\,e$ and $b\,\mathcal{R}\,e$.*

Fact 8 (from Green's lemma). *All $\mathcal{H}$-classes in a $\mathcal{D}$-class have the same cardinality.*

Fact 9. *Let H be an $\mathcal{H}$-class in S. Either for all a, b in H, $ab \notin H$; or for all a, b in H, $ab \in H$, and furthermore $(H, .)$ is a group.*

Case of a group $\mathcal{H}$-class.

Lemma 10. *Let H be an $\mathcal{H}$-class in S such that $(H, .)$ is a group, and β be such that $\sigma(\beta) \subseteq H$. Then there exists a ramseyan split of height at most $|H|$ of (β, σ).*

Proof. Since $(H, .)$ is a group, it is natural to extend the definition of σ over β in the following way. For every x, let $\sigma(x, x)$ be 1_H, the neutral element of the group $(H, .)$; for every $y < x$ in β, let $\sigma(x, y)$ be $\sigma(y, x)^{-1}$, the inverse of $\sigma(x, y)$ in H. As expected, this extended version of σ satisfies for every x, y, z in β, $\sigma(x, z) = \sigma(x, y)\sigma(y, z)$. Let n be a mapping numbering the elements of H from 1 to $|H|$. Fix an element x_0 in β. Let s be defined for all x by $s(x) = n(\sigma(x_0, x))$.

Let us show that s defined this way is indeed a ramseyan split for σ. Let $x < y$ be such that $s(x) = s(y)$, then $\sigma(x_0, x) = \sigma(x_0, y)$ since n is a bijection from H onto $[1, |H|]$. Hence $\sigma(x, y) = \sigma(x, x_0)\sigma(x_0, y) = \sigma(x_0, x)^{-1}\sigma(x_0, y) = 1_H$. Hence, given $x < y$ and $x' < y'$ pairwise k-neighbours, then $\sigma(x, y) = 1_H = \sigma(x', y') = 1_H^2$. □

Case of a regular $\mathcal{D}$-class.

Lemma 11. *Let D be a regular $\mathcal{D}$-class in S, and β be such that $\sigma(\beta) \subseteq D$. Then there exists a ramseyan split of height at most $|D|$ of (β, σ).*

Proof. For every $x \in \beta$ nonmaximal, set $r(x)$ to be the $\mathcal{R}$-class of $\sigma(x, z)$ for some $z > x$; this value is independant of the choice of z according to Fact 7. Similarly, for every x in β nonminimal, set $l(x)$ to be the $\mathcal{L}$-class of $\sigma(y, x)$ for some $y < x$. If β has a maximal element M, choose $r(M)$ to be such that $l(M) \cap r(M)$ is a subgroup of S; this is possible according to Fact 6. Similarly if β has a minimal element m, choose $l(m)$ such that $l(m) \cap r(m)$ is a subgroup of S. Set for all x in β, $h(x) = l(x) \cap r(x)$.

We claim that for every x in β, $h(x)$ is a subgroup of S. Indeed, if x is either the minimal or the maximal element of β, this follows from the definition of $r(M)$ and $l(m)$. Else, there exists y, z such that $y < x < z$. Let a be $\sigma(y, x) \in l(x)$ and b be $\sigma(x, z) \in r(x)$. By Fact 7, since $ab = \sigma(y, z) \in D$, there exists an idempotent e in D such that $a \mathcal{L} e$ and $b \mathcal{R} e$; i.e. $e \in h(x)$. And by Fact 9, $h(x)$ is a subgroup of S. The claim holds.

According to Fact 8, there is a natural number N such that all $\mathcal{H}$-classes included in D have cardinal N. Let $H_1, \ldots, H_d$ be the $\mathcal{H}$-classes included in D which are subgroups of S. For k in $\{1, \ldots, d\}$, set β_k to be $\{x \in \beta \ : \ h(x) = H_k\}$. By fact 7, $\sigma(\beta_k) \subseteq H_k$. By Lemma 10, there exists a ramseyan split s_k for (β_k, σ) of height at most $|H_k| = N$.

We set now for all x in β, $s(x)$ to be $kN + s_k(x)$ where k is such that $x \in \beta_k$. Let us establish that s is a ramseyan split for (β, σ). Let $x < y$ and $x' < y'$ be such that $s(x) = s(y) = s(x') = s(y')$. By definition of s, x, y, x', y' belong to the same β_k. Furthermore, since $s(x) = s(y) = s(x') = s(y')$, we have $s_k(x) = s_k(y) = s_k(x') = s_k(y')$. Hence, by ramseyanity of s_k over (β_k, σ), $\sigma(x, y) = \sigma(x', y') = \sigma(x, y)^2$. We conclude that the mapping s is a ramseyan split for (β, σ). Its height is bounded by $dN \leq |D|$. □

The general case for ordinals: proof of Theorem 3.

For this last part of the proof, one has to provide factorisations on ordinals where the minimal value has ben removed. Without this, one does not obtain the bound of $|S|$ announced. Hence, given a linear well-ordering β, one denotes by $\dot{\beta}$ the linear ordering $\beta \setminus \{0_\beta\}$.

Lemma 12. *Let $E \subseteq S$ be a $\mathcal{D}$-closed subset of S and $\beta \subseteq \alpha$ be such that $\sigma(\beta) \subseteq E$. Then there exists a ramseyan split of height at most $|E|$ of $(\dot{\beta}, \sigma)$.*

Proof. The proof is done by induction on the size of E. If E is empty, then β contains at most one element. Hence $\dot{\beta}$ is empty. We can give a split of height 0 over the empty linear ordering.

Else, let D be a minimal $\mathcal{D}$-class in E (for the $\leq_{\mathcal{J}}$-order). Let $\gamma \subseteq \beta$ be the least set satisfying:

- $0_\beta \in \gamma$, where 0_β is the minimal element of β,
- if $x \in \gamma$ then $\min\{y > x \ : \ \sigma(x, y) \in D\} \in \gamma$.

It is not difficult to check that the following fact holds.

Fact 13. *For every x, y in β, if $]x, y] \cap \gamma$ is empty, then $\sigma(x, y) \notin D$. If $[x, y] \cap \gamma$ contains two elements, then $\sigma(x, y) \in D$.*

Define the equivalence relation $\sim$ over β by $x \sim y$, if $]x, y] \cap \gamma = \emptyset$ for $x < y$ and closed under reflexivity and symmetry. Let η be an equivalence class for $\sim$. By Fact 13, $\sigma(\eta) \cap D = \emptyset$. Hence, one can apply the induction hypothesis and obtain a ramseyan split $s_{\dot{\eta}}$ for $(\dot{\eta}, \sigma)$ of height at most $|E| - |D|$. Remark that $\dot{\eta} = \eta \setminus \gamma$.

At this point, two cases may happen depending on the regularity of D. If D is not regular, then γ contains at most 2 elements, Indeed, assume $x < y < z$ in γ, then $\sigma(x, y), \sigma(y, z)$ and $\sigma(x, y)\sigma(y, z) = \sigma(x, z)$ belong to D. By Fact 6, D would be regular. A contradiction. Define $s_{\dot{\beta}}$ over $\dot{\beta}$ by $s(x) = 1$ for $x \in \gamma$, else $s(x) = s_{\dot{\eta}}(x) + 1$ for η the equivalence class of x. This split is ramseyan since the value 1 is used at most once (in $\dot{\gamma}$), and the ramseyanity is inherited from the induction hypothesis elsewhere. By induction hypothesis, this split has height at most $|E| - |D| + 1 \leq |E|$.

Finally, if D is regular. We have $\sigma(\gamma) \subseteq D$. By Lemma 11 we obtain ramseyan split s_γ of height at most $|D|$ for (γ, σ). Then define s over $\dot{\beta}$ by $s(x) = s_\gamma(x)$ for $x \in \gamma$, else $s(x) = |D| + s_\eta(x)$ for η the equivalence class of x. It follows from the definition that s is a ramseyan split of $(\dot{\beta}, \sigma)$ of height at most $|E| - |D| + |D| = |E|$. □

We can now conclude the proof of Theorem 3.

Proof. Given an ordinal α, and an additive labelling σ from α to S. Fix a value a_0 in S, construct the linear ordering $\alpha' = 1 + \alpha$, where 1 is a linear ordering containing the single element 0. Set $\sigma'(x,y)$ for $x < y$ in α to be $\sigma(x,y)$, and set $\sigma'(0,y)$ to be $a_0.\sigma(0_\alpha, y)$. Defined like this, σ' is an additive labelling from α' to S. By Lemma 12, there exists a ramseyan split s for $(\dot{\alpha}', \sigma')$ of height at most $|S|$. By construction of α' and σ', s is also a ramseyan split for (α, σ). □

The general case for complete orderings: proof of Theorem 4.

Theorem 4 follows directly from the following lemma, with $E = S$.

Lemma 14. *Let $E \subseteq S$ be a $\mathcal{D}$-closed subset of S and $\beta \subseteq \alpha$ be complete and such that $\sigma(\beta) \subseteq E$. Then there exists a ramseyan split of height at most $3|E|$ of (β, σ).*

Proof. We assume wlog that β is nonempty. The proof is done by induction on the size of E. Let D be a minimal $\mathcal{D}$-class in E (for the $\leq_{\mathcal{J}}$-order). We define a binary relation R over β by: for every $x < y$ in β, $R(x,y)$ if $\sigma(x,y) \in D$. Since D is a minimal $\mathcal{D}$-class, this relation is upward closed; we can apply Lemma 1 and obtain a set γ satisfying its conclusion.

Define the equivalence relation $\sim$ over $\beta \setminus \gamma$ by $x \sim y$, if $[x,y] \cap \gamma = \emptyset$ for $x < y$ and closed under reflexivity and symmetry. Let η be an equivalence class for $\sim$. By Lemma 1 from which is obtained γ, $\sigma(\eta) \cap D = \emptyset$. Hence, one can apply the induction hypothesis and obtain a ramseyan split s_η for (η, σ). At this point, two cases may happen depending on the regularity of D.

If D is not regular, then γ contains at most 2 elements (same argument as in the case of α being an ordinal. Let us treat the case of γ containing two elements $x_0 < x_1$ (the case of γ being empty or a singleton can be deduced from it). The equivalence $\sim$ has at most three equivalence classes, $\eta = (-\infty, x_0[$, $\eta' =]x_0, x_1[$, and $\eta'' =]x_1, +\infty)$. We can apply the induction hypothesis with $\sigma(\eta) \subseteq E \setminus D$ (resp. $\sigma(\eta') \subseteq E \setminus D$ and $\sigma(\eta'') \subseteq E \setminus D$) and obtain a ramseyan split s_η for (η, σ) (resp. $s_{\eta'}$ for (η', σ) and $s_{\eta''}$ for (η'', σ)) of height at most $3(|E| - |D|)$. We construct s over β by $s(x) = s_\eta(x) + 2$ if $x \in \eta$, $s(x_0) = 1$, $s(x) = s_{\eta'}(x) + 2$ if $x \in \eta'$, $s(x_1) = 2$, and $s(x) = s_{\eta''}(x) + 2$ for $x \in \eta''$. It follows from the definition that s is a ramseyan split of (β, σ) of height at most $3(|E| - |D|) + 2 \leq 3|E|$.

Else, if D is regular, we apply Lemma 3 on γ with $k = 3$ and obtain a mapping $c : \gamma \to \{0, 1, 2\}$ satisfying the conclusions of Lemma 3. By Lemma 1, $\sigma(c^{-1}(0)) \subseteq D$. We can apply Lemma 11 to $c^{-1}(0)$, obtaining a ramseyan split s' for $(c^{-1}(0), \sigma)$ of height at most $|D|$. Let x be in β, we define

$$s(x) = \begin{cases} s'(x) & \text{if } x \in \gamma, \text{ and } c(x) = 0 \\ |D| + c(x) & \text{if } x \in \gamma, \ c(x) \in \{1, 2\} \\ s_\eta(x) + |D| + 2 & \text{if } x \notin \gamma, \text{ and } \eta \text{ is the } \sim\text{-equivalence class of } x. \end{cases}$$

Let us first remark that the values corresponding to the first case of the definition range in $[1, |D|]$ (def. of s'). The values of the second case lie in $[|D| + 1, |D| + 2]$ by construction. Finally, the values provided by the last case lie all in $[|D| + 3, |D| + 2 + 3(|E| - |D|)]$, which is included in $[|D| + 3, 3|E|]$.

We have to prove the ramseyanity of s. Let $x < y$ and $x' < y'$ be pairwise k-neighbours for some k. If $k \in [1, |D|]$, we are in the first case of the definition of s, and $\sigma(x,y) = \sigma(x',y') = \sigma(x,y)^2$ by ramseyanity of s'. If $k \in [|D| + 1, |D| + 2]$, then $c(x) = c(y)$ and by Lemma 3, there is some z in $]x, y[$ with $c(z) = 0$. This implies $s(z) \leq |D|$, contradicting the 'k-neighbourity' of x and y. Finally if $k \geq |D| + 3$, since x, y, x' and y' are k-neighbours, they all lie in the same $\sim$-equivalence class η. And $\sigma(x,y) = \sigma(x',y') = \sigma(x,y)^2$ by ramseyanity of s_η. □

5 Application to countable scattered linear orderings

In this section, we use Theorem 4 for giving a new simplified proof of Theorem 5 (known from [10]). We first briefly recall some facts about scattered linear orderings in Section 5.1 and define the corresponding notions for words. Then we introduce automata on countable scattered words in Section 5.2 and the corresponding algebraic definition of a $\diamond$-semigroup in Section 5.3. In Section 5.4, we prove Theorem 5.

This section is independant from the subsequent ones.

5.1 Scattered linear orderings

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 it is not dense on any subordering. 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.

Given an alphabet A, we denote by $A^\diamond$ the set of words indexed by a countable scattered linear ordering. Given a language $L \subseteq A^\diamond$, L^ω represents the set of words of the form $\prod\{u_i \ : \ i \in \omega\}$ where all the u_i's belong to L. One defines similarly $L^{-\omega}$ and L^ζ.

A standard way for proving results on scattered linear orderings is to use the theorem of Hausdorff (chapter 5 of [22] is dedicated to the subject). It establishes a general way of decomposing scattered linear orderings. Hausdorff's theorem is a key tool in the original proof of Theorem 5 [10]. We avoid it below; instead, we use the following lemma which provides a kind of induction principle for scattered linear orderings. It essentially says that an equivalence relation such that any two sets of equivalent elements are contiguous (there is nothing in between) are equivalent, then the relation contains is trivial.

Lemma 15. *Given a scattered linear ordering α and an equivalence relation R over α satisfying:*

$$\textit{for all } X < Y, \textit{ with } X^2 \subseteq R, \textit{ and } Y^2 \subseteq R, \qquad \bigcap_{x \in X,\ y \in Y}]x, y[= \emptyset \quad \textit{implies } (X \cup Y)^2 \subseteq R \ ;$$

Then $R = \alpha^2$.

Proof. Consider the set S of equivalence relations included in R such that every equivalence class is convex. It is nonempty since the equality relation over α belongs to S. Order S by inclusion. Given a chain in S, the union of all relations in the chain is itself an element of S: the chain has an upper bound in S. Then, according to Zorn's lemma, there is a maximal element $\sim$ in S. Since α is scattered and $\sim \in S$, $\alpha/_\sim$ is itself a scattered linear ordering. Assume that it has two distinct equivalence classes. Since $\alpha/_\sim$ is scattered, there are two equivalence classes X and Y — choose wlog $X < Y$ — such that there is no other equivalence class Z with $X < Z < Y$. This follows that $\cap_{x \in X,\ y \in Y}]x, y[= \emptyset$. Applying the hypothesis leads to $(X \cup Y)^2 \subseteq R$, and consequently $(\sim \cup (X \cup Y)^2) \in S$. It contradicts the maximality of $\sim$. □

5.2 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 [7], but in their 'Muller' form, while here we adopt the 'parity-like' approach.

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 left limit, then $(k, \rho(c)) \in \delta$ where $k = \max \bigcap_{c'<c} p(\rho(]c', c[))$,
- if c is a right 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 4. 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 apply to the right of c. In both cases there is a contradiction.

The languages accepted by priority automata are closed under union, intersection, concatenation, projection and exponentiation by ω and $-\omega$ [7]. They also admit an equivalent form of regular expressions [7] and their emptyness problem is decidable. A consequence of Theorem 5 below is their closure under complementation (originally proved in [10], in [9] for a particular case).

5.3 On $\diamond$-semigroups

Finite semigroups are known to have the same 'expressive power' as finite state automata. This approach has been extended to languages of ω-words while introducing ω-semigroups in [16]. Then Bedon and Carton generalized it to words indexed by countable ordinals in [2], the corresponding algebraic object being called an ω_1-semigroup. Finally, Carton and Rispal have introduced $\diamond$-semigroups for describing languages of words indexed by scattered linear orderings.

Formally, a $\diamond$-semigroup (s, π) is a set equipped with an operator π mapping $S^\diamond$ to S which satisfies:

- for all $s \in S$, $\pi(s) = s$, and,
- for all countable scattered linear ordering α and families $(u_i)_{i \in \alpha}$ of words in $S^\diamond$,

$$\pi(\prod\{\pi(u_i) \ : \ i \in \alpha\}) = \pi(\prod\{u_i \ : \ i \in \alpha\}) \ .$$

Those properties express the fact that π is a generalized product operator: more precisely, the rules correspond to a generalized form of associativity. For instance, for every u, v, w in S,

$\pi(u\pi(vw)) = \pi(uvw) = \pi(\pi(uv)w)$. In this sense, every $\diamond$-semigroup can be seen as a semigroup with the product defined by $u.v = \pi(uv)$. The free $\diamond$-semigroup generated by a finite alphabet A is $(A^\diamond, \prod)$.

Given two $\diamond$-semigroups (S, π) and (S', π'), a mapping φ from S to S' is a *morphism of $\diamond$-semigroups* if for every scattered linear ordering α, and every $(x_i)_{i\in\alpha}$ in S, $\varphi(\pi(\prod\{x_l \ : \ l \in \alpha\})) = \pi'(\prod\{\varphi(x_l) \ : \ l \in \alpha\})$. A language $K \subseteq A^\diamond$ is *$\diamond$-recognizable* if there exists a morphism of $\diamond$-semigroups from $A^\diamond$ to a finite $\diamond$-semigroup saturating K; i.e. such that $\varphi^{-1}(\varphi(K)) = K$. As usual with recognizability, $\diamond$-recognizable languages are closed under union, intersection and complementation.

From now, we denote $\pi(uv)$ simply by uv. More generally, given a word u in $S^\diamond$, we do not distinguish between u and $\pi(u)$. Similarly, we abbreviate $\pi(\prod\{u \ : \ i \in (\mathbb{N}, <)\})$ by u^ω and $\pi(\prod\{u \ : \ i \in (-\mathbb{N}, <)\})$ by $u^{-\omega}$. We also denote by u^ζ the value $u^{-\omega}u^\omega$.

Example 5. Consider the set $S = (\{0, 1\} \times \{0, 1\}) \uplus \{\bot\}$. Define the product . and the exponent mappings ω and $-\omega$ by, for every x in S and a, b, a', b' in $\{0, 1\}$,

$$\bot x = x\bot = \bot \qquad (a, b)(a', b') = \begin{cases} \bot & \text{if } b = a' = 1 \\ (a, b') & \text{else} \end{cases}$$

$$\bot^\omega = (1, 1)^\omega = \bot \qquad (a, b)^\omega = \begin{cases} \bot & \text{if } a = b = 1 \\ (a, 1) & \text{else} \end{cases}$$

$$\bot^{-\omega} = (1, 1)^{-\omega} = \bot \qquad (a, b)^{-\omega} = \begin{cases} \bot & \text{if } a = b = 1 \\ (1, b) & \text{else.} \end{cases}$$

Using Theorem 10 in [10], this $(S, .)$ together with the mappings ω and $-\omega$ defines uniquely a $\diamond$-semigroup (S, π).

Let u be in $\{a\}^\diamond$ of domain α. Set $\varphi(u)$ to be $\bot$ if α is not complete. If α is complete, set $\varphi(u)$ to be (a, b) where $a = 0$ if α has a minimal element, else $a = 1$, and $b = 0$ if α has a maximal element, else $b = 1$. This φ is a morphism from $(\{a\}^\diamond, \prod)$ to (S, π). It follows that the set of words in $\{a\}^\diamond$ of complete domain is $\diamond$-recognizable: it is equal to $\varphi^{-1}(\{0, 1\} \times \{0, 1\})$.

5.4 Equivalence of representations

The following theorem was proved in [10][2]. A direct consequence of it is the closure under complementation of the languages of words indexed by scattered linear orderings accepted by priority automata.

Theorem 5 ([10]). *Let A be a finite alphabet. A language $L \subseteq A^\diamond$ is accepted by a priority automaton if and only if it is $\diamond$-recognizable.*

The left to right implication is standard: one constructs a $\diamond$-semigroup which captures all the possible behaviours of the automata over a word. Then there is no choice on the definition of the product and the morphism.

The difficult direction is, given a $\diamond$-recognizable language, to construct a priority automaton accepting it. The contribution here is to show that a natural way of constructing such an automaton is to follow the structure of a ramseyan split. Let us fix a $\diamond$-semigroup (S, π) and a morphism of $\diamond$-semigroups φ from $(A^\diamond, \prod)$ to (S, π). By closure of priority automata under union, it is sufficient to show that for every $c \in S$ the language $\varphi^{-1}(c)$ is acepted by a priority automaton.

[2] In fact, the present theorem differs in the use of priority automata in place of automata using Muller condition in limit transitions. For this reason the result here is new; but for a nonessential reason.

Let k be a natural number, set L_k to be the set of words u such that φ_u admits a ramseyan split of height at most k. We show by induction on k that for every $c \in S$, the language $L_{c,k} = L_k \cap \varphi^{-1}(c)$ is accepted by an automaton. According to Theorem 4 we have $\varphi^{-1}(c) = L_{c,3|S|}$. We also use the intermediate language $SD(e,k)$ for e an idempotent of S which is the set of words u of domain α admitting a ramseyan split s of height at most k, such that $s(\bot_\alpha) = s(\top_\alpha) = 1$ and $\varphi(u) = e$ (in particular, $SD(e,k) \subseteq L_{e,k}$).

The following lemma reduces the problem from describing the language $L_{c,k}$ to describing languages of the form $SD(e,k)$.

Lemma 16. *Let $u \in A^\diamond$ be a word of at least two letters. Then u belongs to $L_{c,k+1}$ iff there exists a, b, e in S and $\gamma \in \{0, 1, \omega, -\omega, \zeta\}$ such that $e^2 = e$, $c = ae^\gamma b$ and $u \in L_{a,k}(SD(e,k+1))^\gamma L_{b,k}$ (with the convention that $xy^0z = xz$).*

Proof. From left to right. Let u be an α-word in $A^\diamond$ of length at least 2, and let s be a ramseyan split of height at most $k+1$ of $(\overline{\alpha}, \varphi_u)$. We argue on the nature of $s^{-1}(1)$.

If $s^{-1}(1)$ is empty, then choose arbitrarily a cut c in $\overline{\alpha}^*$, and set a new value of 1 to $s(c)$. This modified s is still a ramseyan split of height $k+1$ of $(\overline{\alpha}, \varphi_u)$. And we can apply the next case for which $s^{-1}(1)$ is a singleton.

If $s^{-1}(1)$ is a singleton $\{c\}$, let v be u restricted to positions to the left of c, and w be u restricted to positions to the right of c. Obviously $u = vw$, and we have $u \in L_{\varphi(v),k} e^0 L_{\varphi(w),k}$ for any idempotent e.

Else $s^{-1}(1)$ contains at least two elements. There are four cases depending on the existence of a minimal (resp. a maximal) element in $s^{-1}(1)$. First case. If $s^{-1}(1)$ has both a minimal element c and a maximal element c', then let $a = \varphi_u(\bot, c)$, $e = \varphi_u(c, c')$, and $b = \varphi_u(c', \top)$. By definition of a ramseyan split, e is an idempotent of S; furthermore, $\varphi(u) = aec$. We obtain $u \in L_{a,k} SD(e, k+1) L_{b,k}$. Second case. If $s^{-1}(1)$ has neither a minimal element nor a maximal element. Let c be $\inf(s^{-1}(1))$ and c' be $\sup(s^{-1}(1))$. Let $a = \varphi_u(\bot, c)$, $b = \varphi_u(c', \top)$. Using the countability of $\overline{\alpha}^*$, we have a ζ-indexed sequence $\cdots < x_n < x_{n+1} < \cdots$ in $s^{-1}(1)$, such that $\inf\{x_i \ : \ i \in \zeta\}$ is c, and $\sup\{x_i \ : \ i \in \zeta\}$ is c'. Let e be $\varphi_u(x_1, x_2)$. The sequence of x_i's shows that $\varphi_u(c, c') \in (SD(e, k+1))^\zeta$. Furthermore e is an idempotent. We obtain $u \in L_{a,k}(SD(e,k+1))^\zeta L_{b,k}$. The two other cases are obtained as combinations of the two first one, using ω and $-\omega$-indexed sequences. □

This lemma together with the closure properties of languages accepted by priority automata shows that it is sufficient to construct an automaton accepting $SD(e, k+1)$. For this, define the following languages:

$$M_{e,k} = \{u \in L_k \setminus \{\varepsilon\} : \ \varphi(u) = e\}, \qquad M_{e,k}^{\leftarrow} = \{u \in L_k \ : \ \varphi(u)e^{-\omega} = e\},$$
$$M_{e,k}^{\rightarrow\leftarrow} = \{u \in L_k \ : \ e^\omega \varphi(u) e^{-\omega} = e\}, \qquad M_{e.k}^{\rightarrow} = \{u \in L_k \ : \ e^\omega \varphi(u) = e\}.$$

Those languages can be obtained as unions of the $L_{a,k}$ together with languages consisting of a single letter word, or the empty word. Hence, by induction hypothesis there are automata accepting them. We identify below the automaton and the language.

In order to accept the language $SD(e, k+1)$, we construct a corresponding automaton $\mathcal{A}(e, k+1)$. The definition of the automaton $\mathcal{A}(e, k+1)$ is depicted in Figure 2. This is a disjoint union of the automata accepting $M_{e,k}, M_{e,k}^{\leftarrow}, M_{e,k}^{\rightarrow}$ and $M_{e,k}^{\rightarrow\leftarrow}$ and of a new state t of priority n; the state t being both initial and final. The value n is chosen to be the highest priority of the automaton. New ε-transitions[3] are added to this construction as depicted in Figure 2: arrow arriving from the left have the initial states of the automaton as destination, while the arrows leaving to the

[3] ε-transitions are just a commodity notation. And in particular there is no cycle of such transitions.

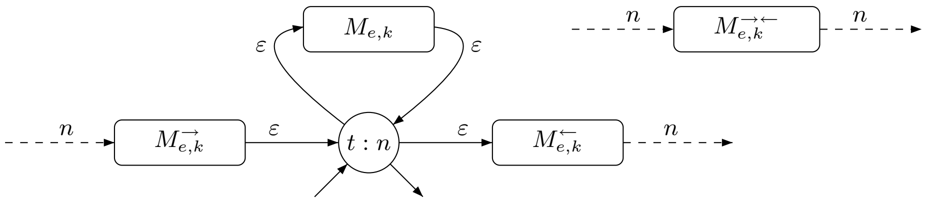

Fig. 2. The automata $\mathcal{A}(e, k+1)$

right have the final states of the automaton as origin. Dashed arrows represent limit transitions. For instance the leftmost one expresses the existence of a limit transition (n, q) for q an initial state of $M^{\rightarrow}_{e,k}$: the automaton can go to state q if the maximal priority appearing infinitely often to its left is n. The following lemma concludes the proof.

Lemma 17. *The automaton $\mathcal{A}(e, k+1)$ accepts the language $SD(e, k+1)$.*

Proof. From right to left. Let u be a word indexed by α. Let s be a ramseyan split of φ_u corresponding to the membership of u in $SD(e, k+1)$, i.e. such that $s(\bot) = s(\top) = 1$.

We construct a run $\rho \in Q^{\overline{\alpha}}$ in the following way (Q is the set of states of $A_{e,k+1}$). Set $\rho(x) = t$ whenever $s(x) = 1$. We define ρ elsewhere by copying runs of the automata $E_{e,k}, M^{\leftarrow}_{e,k}, M^{\rightarrow}_{e,k}$ and $M^{\rightarrow\leftarrow}_{e,k}$. More precisely, consider a maximal interval $I \subseteq \overline{\alpha}$ such that $s(I) \geq 2$. Let us define ρ over I. Four cases happen depending on the nature of the interval: $I = [x, y], [x, y[,]x, y]$ or $]x, y[$. We treat the case of $[x, y[$. The others are similar.

If $I = [x, y[$, this means that $s(x) > 1$, but $s(y) = 1$. As a consequence, there is a sequence $x_1 < x_2 < \dots$ in $s^{-1}(1)$ indexed by ω such that $\sup\{x_i : i < \omega\} = x$ (this is possible because $\overline{\alpha}$ is countable). It follows that $\sigma(x_1, x) = e^\omega$. Furthermore (by ramseyanity), $\sigma(x_1, y) = e$. We deduce $e^\omega \sigma(x, y) = e$. By induction hypothesis, we obtain that v is accepted by $M^{\rightarrow}_{e,k}$. We define ρ to replicate the corresponding run over I using the instance of $M^{\rightarrow}_{e,k}$ it contains. We have to prove that this choice indeed produces a run. Over $]x, y[$ this is a correct run since the original run was itself correct. It remains to show the correctness of the run to the left of x. But, we already know that the maximal priority reaching x from the left is n since the sequence of the x_i's tends to x and by construction correspond to a priority n which is maximal. We conclude that there is a corresponding transition in $A_{e,k+1}$.

From left ro right. Let $\rho \in Q^{\overline{\alpha}}$ be a run of $A(e, k+1)$ over u from t to t. We aim at constructing a ramseyan split s of φ_u corresponding to the membership of u in $SD(e, k+1)$. Let J be $\rho^{-1}(t)$. We set $s(x)$ to be 1 over J. Let I be a maximal interval which does not intersect J. Once more there are four cases: $I = [x, y], [x, y[,]x, y]$ or $]x, y[$. We treat the case of $[x, y[$. The others being similar.

If $I = [x, y[$, this means that $s(x) > 1$, but $s(x) = 1$. Let q be the state $\rho(x)$. Since I is maximal, there exists an ω-sequence $x_1 < x_2 < \dots$ in J of limit x. Since $\rho(x_i)$ is n by definition, this means that the maximal priority appearing infinitely often to the left of x is n. Hence, there must be in $A_{e,k+1}$ a limit transition from n to q. By inspecting the definition of $A_{e,k+1}$, this means that q is either the initial state of $M^{\rightarrow}_{e,k}$ or the initial state of $M^{\rightarrow\leftarrow}_{e,k}$. In y, the run assumes state n, but this state has been reached by an ε-transition either from the final state of $M_{e,k}$, or by the final state of $M^{\rightarrow}_{e,k}$. Let p' be this state. We know that there is a run of $A_{e,k+1}$ from configuration (x, q) to (y, p') which does not visit state n (by definition of I). It follows that q is the initial state of $M^{\rightarrow}_{e,k-1}$, p' is its final state and that the run from (x, q) to (y, p') is an accepting run of $M^{\rightarrow}_{e,k}$. By induction hypothesis, $\sigma_{(u|_I)} = (\varphi_u)|_I$ has factorisation height at most k. Let s' be this factorisation. For all $x \in I$, let $s(x)$ be $s'(x) + 1$.

Let us show that this split is ramseyan. Let $x < y$ be such that $s(x) = s(y) = k$. For k-neighours with $k \geq 2$, this is inherited from the induction hypothesis. What remains to be shown is that for every $x < y$ in J (i.e. x, y are 1-neighbours), $\sigma(x, y) = e$. To make this relation reflexive and symmetric, we consider the relation R defined by xRy if $x = y$ or $x < y$ and $\sigma(x, y) = e$ or $y < x$ and $\sigma(y, x) = e$. We want to apply Lemma 15 on $(J, <)$ and the relation R. Let $X, Y \subseteq J$ be such that $X < Y$, $X^2 \subseteq R$, $Y^2 \subseteq R$ and $\cap_{x \in X, y \in Y}]x, y[\cap X = \emptyset$. Let $I = \cap_{x \in X, y \in Y}]x, y[$, I is a maximal interval nonintersecting J.

Once more there are four cases: $I = [x, y], [x, y[,]x, y]$ or $]x, y[$. We treat the case of $I = [x, y[$. Fix $x_0 \in X$ and $y_0 \in Y$. We want to prove $\sigma(x_0, y_0) = e$. As $x \notin J$, there is an ω-sequence $x_0 < x_1 < \ldots$ of limit x with for all i, $\sigma(x_i, x_{i+1}) = e$. It follows that $\sigma(x_0, x) = e^\omega$. By construction s corresponds to a run of $M_{e,k}^{\rightarrow}$ over I. It follows, by definition of $M_{e,k}^{\rightarrow}$, that $e^\omega \sigma(x, y) = e$. We obtain $\sigma(x_0, y) = e$. Since furthermore by hypothesis, $\sigma(y, y_0) = e$, we have $\sigma(x_0, y_0) = e$.

Lemma 15 concludes that for every $x < y$ in J, $\sigma(x, y) = e$. Hence, s is a ramseyan split for φ_u. □

6 Deterministic extension to the factorisation forest theorem

We try in this section to construct the split from 'left to right' in a 'deterministic way'. The notion of ramseyanity is not suitable anymore in this context; the result would be false[4]. It is replaced by the notion of forward ramseyanity. The result, Theorem 6, only holds for ordinals.

6.1 The statement

A split s of height N is *forward ramseyan* if for every $k = 1 \ldots n$ and k-neighbours $x < y$ and $x' < y'$,

$$\sigma(x, y) = \sigma(x, y).\sigma(x', y') \ .$$

So in particular, $\sigma(x, y)$ is an idempotent, but $\sigma(x, y)$ and $\sigma(x', y')$ may be different idempotents. In the terminology of Green's relation, $\sigma(x, y)$ and $\sigma(x', y')$ are $\mathcal{L}$-equivalent idempotents. A ramseyan split is always forward ramseyan, but the converse does not hold in general.

Below, we also identify the natural numbers with the corresponding ordinal. Furthermore, for σ an additive labelling over an ordinal α, and given $\beta < \alpha$, we denote by $\sigma|_{\leq\beta}$ the labelling σ restricted to $[0, \beta]$.

Theorem 6. *Let $(S, .)$ be a semigroup. To every additive labelling σ over an ordinal α, one can associate a forward ramseyan split $s_{\alpha,\sigma}$ of (α, σ) of height at most $|S|$. Furthermore, for every additive labellings σ and σ' over the respective ordinals α and α', and every ordinal $\beta < \min\{\alpha, \alpha'\}$,*

$$\textit{if} \quad \sigma|_{\leq\beta} = \sigma'|_{\leq\beta} \quad \textit{then } s_{\alpha,\sigma}(\beta) = s_{\alpha',\sigma'}(\beta) \qquad \textit{(determinism property)} \ .$$

Furthermore, under the same hypothesis, over finite linear orderings, the forward ramseyan split can be computed via monadic formulæ.

Proposition 1 (definable variant of Theorem 6). *Given a finite semigroup $(S, .)$, there exist monadic closed formulæ $\Theta_1, \ldots, \Theta_{|S|}$ such that for every ordinal α, and additive labelling σ from α to S, the split s defined for every $\beta \in \alpha$ by:*

$$s(\beta) = n \quad \textit{such that} \quad (\beta + 1, \sigma|_{\leq\beta}) \models \Theta_n \ ,$$

is forward ramseyan.

[4] Consider the semigroup $(\{a, b\}, .)$ defined by $ab = aa = a$ and $ba = bb = b$.

Proof. (Idea) Implement the construction of the proof of Theorem 6 via monadic formulæ. □

Note, that a consequence of Proposition 1, the mapping s satisfies the determinism property.

6.2 Proof of Theorem 6

Once more, we perform a case analysis.

Case of a single $\mathcal{H}$-class.

Lemma 18. *Let H be an $\mathcal{H}$-class in S such that $(H, .)$ is a group. For $\beta \subseteq \alpha$ such that $\sigma(\beta) \subseteq H$, there exists a ramseyan split $s^H_{\beta,\sigma}$ of height at most $|H|$. Furthermore s^H satisfies the determinism property.*

This is exactly the proof of Lemma 10 in which one always chooses x_0 to be 0_β.

Case of a single $\mathcal{L}$-class.

Lemma 19. *Let L be an $\mathcal{L}$-class in a regular $\mathcal{D}$-class, for every ordinal β such that $\sigma(\beta) \subseteq L$, there exists a ramseyan split $s^L_{\beta,\sigma}$ of height at most $|L|$. Furthermore s^L satisfies the determinism property.*

We require the following result.

Fact 20. *There is an $\mathcal{H}$-class $H \subseteq L$ which is a group, and a mapping $f : L \to H$ such that:*

- *for every a, b in L, if $ab \in L$ then $f(ab) = f(a)f(b)$, and,*
- *for every $\mathcal{H}$-class $H' \subseteq L$, $f|_{H'}$ is a bijection from H' onto H.*

Proof. Let $H_1, \dots, H_n$ be the $\mathcal{H}$-classes included in L. By Fact 9 we can assume that $H_1, \dots, H_k$ are groups, while for every a, b in K_i for $i > k$, $ab \notin L$. By regularity hypothesis and Fact 6, $k \geq 1$. Let $L' = H_1 \cup \dots \cup H_k$.

Let a, b be in L, we claim that $ab \in L$ iff $b \in L'$. Indeed, if $b \in L'$, let e be the neutral element of the group containing b. Since $e \mathrel{\mathcal{L}} a$, $e = xa$ for some x. Hence, $b = eb = xab$, and we deduce $ab \mathrel{\mathcal{L}} b$. Conversely, suppose ab in L, then $ab \mathrel{\mathcal{R}} a$. Hence, $a = abc$ for some c. But then $abcbc = a$. Hence bc belongs to L'. But $bc \mathrel{\mathcal{R}} b$. Hence $b \in L'$.

Let H be H_1. If $k = 0$, then for all a, b in L, $ab \notin L$. One can construct the mapping arbitrarily using Fact 8. Else, let e_i be the neutral element of H_i for $i \leq k$. Let $i, j \leq k$. Since $e_i \mathrel{\mathcal{L}} e_j$, $e_i = xe_j$ for some x. Hence $e_ie_j = xe_je_j = xe_j = e_i$. For every $a \in L$, let $f(a) = ae_1$. According to the claim above, f is a mapping from L to H_1. Assume a, b in L such that $ab \in L$. According to the claim, above, $b \in L'$, i.e. $b \in H_i$ for $i \leq k$. Also, as $a \mathrel{\mathcal{L}} e_i$, $a = xe_i$ for some x. We have $f(a)f(b) = ae_1be_1 = xe_ie_1be_1 = xe_ibe_1 = abe_1 = f(ab)$.

The fact that $f|_{H_i}$ is a bijection from H_i to H_1 is known as Green's lemma. □

We can now prove Lemma 19.

Proof. Let H and f be obtained by Fact 20. For $x < y$ in β, let $\sigma'(x, y)$ be $f(\sigma(x, y))$. The first property of f makes σ' an additive labelling from β to H, such that $\sigma(\beta) \subseteq H$. Applying the case of a single $\mathcal{H}$-class above we obtain a split $s^H_{\beta,\sigma'}$ forward ramseyan for (β, σ'). There are two different cases.

Either all the $\mathcal{H}$-classes are groups. In this case, one sets $s^L_{\beta,\sigma}$ to be $s^H_{\beta,\sigma'}$. Let us show that s^L is forward ramseyan. Indeed, consider $x < y$ and $x' < y'$ to be k-neighbours for some k. This means that $f(\sigma(x, y))$ and $f(\sigma(x', y'))$ are equal to the neutral element 1 of H. Since the H-class of $\sigma(x, y)$ (*resp.* of $\sigma(x', y')$) are groups isomorphic to H, we have that $\sigma(x, y)$ and $\sigma(x', y')$

are idempotents of S. Since $\sigma(x,y) \mathrel{\mathcal{L}} \sigma(x',y')$, $\sigma(x,y) = a\sigma(x',y')$ for some $a \in S$. Hence, $\sigma(x,y)\sigma(x',y') = a\sigma(x',y')^2 = a\sigma(x',y') = \sigma(x,y)$.

Else, if there exists a non-regular $\mathcal{H}$-class in L. This means that L contains at least two $\mathcal{H}$-classes. Define $s^L_{\beta,\sigma}(0_\beta) = 1$, and $s^L_{\beta,\sigma} = s^H_{\beta,\sigma'}(x) + 1$ elsewhere. The split s^L defined this way is forward ramseyan for (β,σ) as above. It has height at most $|H| + 1 \leq 2|H| \leq |L|$.

And this construction satisfies the determinism property. □

Case of a single $\mathcal{D}$-class.

Lemma 21. *Let D be a regular $\mathcal{D}$-class. For every ordinal β such that $\sigma(\beta) \subseteq D$, there exists a ramseyan split $s^D_{\beta,\sigma}$ of height at most $|D|$. Furthermore s^D satisfies the determinism property.*

Proof. We prove the property for every $\mathcal{L}$-closed $E \subseteq D$. This is done by induction on the cardinal of E. If E is an $\mathcal{L}$-class, Lemma 19 concludes.

Else, let L be an $\mathcal{L}$-class in E. Let $\gamma = \{0_\beta\} \cup \{x \in \dot{\beta} \ : \ \sigma(0_\beta, x) \in E \setminus L\}$. By Fact 7, for every $x < y$ in γ, $\sigma(x,y) \in E \setminus L$. On can apply the induction hypothesis, and obtain a split $s^{E\setminus L}$ which is forward ramseyan for (γ,σ) and of height at most $|E| - |L|$. Similarly, for every $x < y$ in $\beta\setminus\gamma$, $\sigma(x,y) \in L$. By Lemma 19, one obtains a split s^L which is forward ramseyan for $(\beta\setminus\gamma,\sigma)$ of height at most $|L|$. Let us define the split s^E by $s^E(x) = s^{E\setminus L}(x) + |L|$ if $x \in \gamma$, else $s^E(x) = s^L(x)$ if $x \in \beta\setminus\gamma$. The mapping s^E is forward ramseyan for (β, E) as an inheritance of the forward ramseyanity of s^L and $s^{E\setminus L}$. It has height at most $|E| - |D| + |D| = |E|$. □

For the proof of Theorem 6, we use Lemma 21 with $E = D$, and the same trick as for ordinal ramseyan splits.

7 Compaction of additive labellings

A labelling maps pairs of elements to a finite set (the semigroup): it is defined via a finite number of binary predicates. In this section we show that the use of (forward) ramseyan factorisations permits to encode all this information into a finite number of unary predicates. Furthermore, we show that the whole additive labelling can be reconstructed from those unary predicates via first-order formulæ. We call this technique *compaction*.

As above, there are two variants to the technique. One which usable over complete linear orderings (Section 7.1), and one usable over ordinals, which satisfies furthermore the determinisism property (Section 7.2). In Section 7.3, we apply this technique for proving a new result on monadic interpretations applied to trees. And in Section 7.4 we briefly describe how this result impacts on the theory of finitely presentable infinite structures.

7.1 Compactions of additive labelling over complete linear orderings

We prove here the following statement.

Theorem 7. *For every finite semigroup $(S,.)$ and a in S, there exists a first-order formula $\mathbf{labelling}_a(x,y)$ of free variables x, y, which uses the ordering relation $<$ and unary predicates $p_1,\dots,p_N$ with $N = \lceil(6|S|+2)\log_2(|S|)\rceil$ such that the following holds*[5].

For every complete linear ordering α and additive labelling σ from α to S, there exists subsets $X_1,\dots,X_N$ of α such that for all a in S and $x < y$ in α:

$$\sigma(x,y) = a \qquad \textit{iff} \qquad (\alpha, X_1,\dots,X_N) \models \mathbf{labelling}_a(x,y) \ ,$$

in which for every $i = 1\dots N$, p_i is interpreted as X_i.

[5] We did not try to optimize the value of N.

In this proof, we define first the value of $X_1, \ldots, X_N$, before giving the formulæ.

Using Theorem 3, one obains a ramseyan split s for (α, σ) of height at most $3|S|$. To every element x in α and k with $1 \leq k \leq 3|S|$, we furthermore attach some pieces of information concerning the value of σ. For every k with $1 \leq k \leq 3|S|$, there are two such informations, $l_k(x)$ and $r_k(c)$, taking value in S, and corresponding to a compaction of what is happening to the left of x, and to the right of x respectively. We give the definition of $l_k(x)$, the case of $r_k(x)$ being symmetric.

$$l_k(x) = \begin{cases} \text{any value} & \text{if } L_k(x) = \emptyset \\ \sigma(z, x) & \text{if } L_k(x) \text{ has a maximum } z \\ a & \text{else, with } a \text{ such that } \forall y \in L_k(x). \exists z \in L_k(x).\ z > y \wedge \sigma(z, x) = a \end{cases}$$
$$\text{where } L_k(x) = \{y < x \ : \ s(y) = k\}$$

Note that a consequence of this definition is that, whenver $x < y$ are k-neighbours, then $\sigma(x, y) = l_{s(y)}(y)$. Finally, it is simple to establish that $N = \lceil (6|S| + 2) \log_2(|S|) \rceil$ bits are sufficient for coding $(s(x), l_1(x), \ldots, l_{s(x)}(x), r_1(x), \ldots, r_{s(x)}(x))$.

We have now to construct first-order formulæ which reconstruct the value of $\sigma(x, y)$ for every $x < y$ in α. We do not provide the formulæ explicitly, but instead describe functions which can be easily translated into first-order logic. Let us treat first the 'ascending case'; i.e. compute $\sigma(x, y)$ for $x < y$, $s(x) \leq s(y)$, and $s(z) \geq s(x)$ for all z in $[x, y]$.

Lemma 22. *For every $x < y$ in α, if $s(x) \leq s(y)$ and $s(z) \geq s(x)$ for all z in $[x, y]$, then $\sigma(x, y) = \mathbf{asc}(x, y)$ with:*

$$\mathbf{asc}(x, y) = \begin{cases} l_{s(x)}(y) & \textit{if } s(z) > s(x) \textit{ for all } z \in]x, y[\ , \\ l_{s(x)}(z) l_{s(x)}(y) & \textit{else for some } z \in]x, y[\textit{ with } s(z) = s(x)\ . \end{cases}$$

Proof. Two cases can happen. If for all z in $]x, y[$, $s(z) > s(x)$. This means that $[x, y[\cap s^{-1}(s(x)) = \{x\}$. Hence, by definition, $l_{s(x)}(y) = \sigma(x, y)$.

Else, there exists x' be in $]x, y[\cap s^{-1}(s(x))$. By definition of $l_{s(x)}(y)$, there exists y' in $[x', y[\cap s^{-1}(s(x))$ such that $l_{s(x)}(y) = \sigma(y', y)$. Let now z be the one used in the definition of $\mathbf{asc}(x, y)$. By definition of $l_{s(x)}(z)$, there exists z' in $[x, z[\cap s^{-1}(s(x))$ such that $\sigma(z', z) = l_{s(x)}(z)$. Finally using the ramseyanity of s, we deduce $\sigma(x, y') = \sigma(z', z) = l_{s(x)}(z)$. Overall $\sigma(x, y) = \sigma(x, y')\sigma(y', y) = l_{s(x)}(z) l_{s(x)}(y) = \mathbf{asc}(x, y)$. □

Naturally, there is a corresponding definition for $\mathbf{desc}(x, y)$ satisfying $\sigma(x, y) = \mathbf{desc}(x, y)$ whenever $s(x) \geq s(y)$ and $s(z) \geq s(y)$ for all z in $[x, y]$. Combining $\mathbf{asc}$ and $\mathbf{desc}$ we obtain the following.

Lemma 23. *For every $x < y$ in α, $\sigma(x, y) = \mathbf{labelling}(x, y)$ with:*

$$\mathbf{labelling}(x, y) = \begin{cases} \mathbf{asc}(x, y) & \textit{if } s(x) \leq s(y) \textit{ and } s(z) \geq s(x) \textit{ for all } z \textit{ in } [x, y] \\ \mathbf{desc}(x, y) & \textit{if } s(x) > s(y) \textit{ and } s(z) \geq s(y) \textit{ for all } z \textit{ in } [x, y] \\ \mathbf{desc}(x, z)\mathbf{asc}(z, y) & \textit{else, for } z \in]x, y[\textit{ and } s(z') \geq s(z) \textit{ for all } z' \in [x, y]\ . \end{cases}$$

Proof. There are three cases, corresponding to the three items of the definition. The two first one are treated by Lemma 22 and its variant for $\mathbf{desc}(x, y)$. In the third case, one finds z in $]x, y[$ such that $s(z)$ is minimum. We use Lemma 22 between x and z, and its variant for $\mathbf{desc}$ between z and y, as well as the additivity of the labelling σ, for obtaining:

$$\sigma(x, y) = \sigma(x, z)\sigma(z, y) = \mathbf{asc}(x, z)\mathbf{desc}(z, y) = \mathbf{labelling}(x, y)\ .$$

□

It is not difficult at this point to check that the definition of **labelling** can be translated for every a in S into a first-order formula $\mathbf{labelling}_a$ using as predicate the ordering relation $<$ as well as unary predicates $p_1, \ldots, p_N$ encoding the value of $(s(x), l_1(x), \ldots, l_{s(x)}(x), r_1(x), \ldots, r_{s(x)}(x))$, and satisfying the conclusion of Theorem 7.

7.2 Deterministic compaction of additive labellings over ordinals

We now state a result similar to Theorem 7 in the ordinal case, which satisfies a form of determinism property. The statement in itself is difficult to process; it is similar to the statement of Theorem 7, in which the determinism feature has been injected.

Theorem 8. *For every finite semigroup $(S, .)$ and a in S, there exists a first-order formula $\mathbf{labelling}_a(x, y)$ of free variables x, y, which uses the ordering relation $<$ and unary predicates $p_1, \ldots, p_N$ with $N = \lceil (2|S| + 1) \log_2(|S|) \rceil$ such that the following holds. For every ordinal α and additive labelling σ from α to S, there exists subsets $X_1(\alpha, \sigma), \ldots, X_N(\alpha, \sigma)$ of α such that for all a in S and $x < y$ in α:*

$$\sigma(x, y) = a \qquad \textit{iff} \qquad (\alpha, X_1(\alpha, \sigma), \ldots, X_N(\alpha, \sigma)) \models \mathbf{labelling}_a(x, y) \ ,$$

in which for every $i = 1 \ldots N$, p_i is interpreted as $X_i(\alpha, \sigma)$.

Furthermore, for every additive labellings σ and σ' over the respective ordinals α and α', and every ordinal $\beta < \min(\alpha, \alpha')$,

$$\textit{if } \sigma|_\beta = \sigma'|_\beta \quad \textit{then for all } i, \quad \beta \in X_i(\alpha, \sigma) \textit{ iff } \beta \in X_i(\alpha, \sigma') \qquad \textit{(determinism property)} \ .$$

Let s be the forward ramseyan split of (α, σ) of height $|S|$ obtained by Theorem 6. Let us define $l_k(x)$ as in the previous section (this time only for every $k = 1 \ldots |S|$). Without loss of generality, we assume that there exists a neutral element — denote it 1 — in S, and we set for every x, $\sigma(x, x) = 1$. Define:

$$\mathbf{labelling}(x, y) = \mathbf{labelling}^1(x, y) \ ,$$

with $\mathbf{labelling}^n$ defined by induction for all $n = 1, \ldots, |S| + 1$ by:

$$\mathbf{labelling}^n(x, y) = \begin{cases} 1 & \text{if } n = |S| + 1 \ , \\ \mathbf{labelling}^{n+1}(x, y) & \text{else if } [x, y[\cap s^{-1}(n) = \emptyset \ , \\ \mathbf{labelling}^{n+1}(x, z) l_n(y) & \text{else if } [x, y[\cap s^{-1}(n) = \{z\} \ , \\ \mathbf{labelling}^{n+1}(x, z_0) l_n(z_1) l_n(y) & \text{else if } [x, y[\cap s^{-1}(n) = \{z_0 < z_1 < \ldots\} \ . \end{cases}$$

In this definition, we abbreviate by $[x, y[\cap s^{-1}(n) = \{z_0 < z_1 < \ldots\}$ the fact that z_0 is the minimal element, and z_1 the minimal element different from z_0 in $[x, y[\cap s^{-1}(n)$. Those two elements exist since α is an ordinal and since the case of $[x, y[\cap s^{-1}(n)$ being the emptyset or a singleton is treated above.

The correctness is then stated by the following lemma.

Lemma 24. *For every $x < y$ in α, and $n = 1, \ldots, |S| + 1$, if for all z in $[x, y[$, $s(z) \geq n$, then*

$$\mathbf{labelling}^n(x, y) = \sigma(x, y) \ .$$

Proof. The proof is done by a downward induction on n. For $n = |S|+1$, no z does satisfy $s(z) \geq n$, hence $[x, y[$ has to be empty. It follows that $x = y$, and by consequence $\mathbf{labelling}^n(x, y) = 1 = \sigma(x, y)$.

Else, let $n \leq |S|$. Assume the property true for $n{+}1$ and consider $x \leq y$. Let E be $[x,y[\cap s^{-1}(n)$. If E is empty, this means that for all z in $[x,y[$, $s(z) \geq n+1$. And by induction hypothesis $\mathbf{labelling}^{n+1}(x,y) = \sigma(x,y)$. Hence, $\mathbf{labelling}^{n}(x,y) = \sigma(x,y)$. If E is the singleton $\{z\}$. This means that $l_n(y) = \sigma(z,y)$. It follows that $\mathbf{labelling}^{n}(x,y) = \mathbf{labelling}^{n+1}(x,z)l_n(y) = \sigma(x,z)\sigma(z,y) = \sigma(x,y)$. Finaly, if $E = \{z_0 < z_1 < \dots\}$. By definition of $l_n(z_1)$, $l_n(z_1) = \sigma(z_0,z_1)$. By induction hypothesis, $\mathbf{labelling}^{n+1}(x,z_0) = \sigma(x,z_0)$. Furthermore, by definition of $l_n(y)$, there is some $z > z_0$ such that $l_n(y) = \sigma(z,y)$. Alltogether with the forward ramseyanity of s leads to:

$$\begin{aligned}\mathbf{labelling}_n(x,y) &= \mathbf{labelling}^{n+1}(x,z_0)l_n(z_1)l_n(y)\\ &= \sigma(x,z_0)\sigma(z_0,z_1)\sigma(z,y)\\ &= \sigma(x,z_0)\sigma(z_0,z)\sigma(z,y)\\ &= \sigma(x,y)\ .\end{aligned}$$

□

As in the previous case, the construction is easily adaptable into a presentation by first-order formulæ using the relation $<$ together with $N = \lceil(2|S|+1)\log_2(|S|)\rceil$ unary predicates coding all the possible values of $(s(x), l_1(x), \dots, l_{|S|}(x))$. This concludes the proof of Theorem 8.

7.3 Application to interpretations

We prove in this section Theorem 9. Let us first give two lemmas which are consequences of standard techniques; either the compositional method, or tree automata.

Lemma 25. *Every monadic formula $\Phi(x_1,\dots,x_n)$ is equivalent on trees to a formula of the form $\exists z_1 \dots \exists z_k.\Phi'$ where Φ' is a boolean combination of monadic formulæ of the form $x \sqsubset y \wedge \Psi(x,y)$ (of free variables x,y), $\Psi(x)$ (of free variable x) and $x = y$, for x,y ranging in $\{x_1,\dots,x_n,z_1,\dots,z_k\}$.*

Lemma 26. *For every monadic formula of the form $x \sqsubset y \wedge \Phi(x,y)$ of free variables x,y, there exists a semigroup S_Φ and $A_\Phi \subseteq S_\Phi$ such that, for every tree t, there exists a mapping σ which to every nodes $x \sqsubset y$ associates $\sigma(x,y) \in S_\Phi$, such that*

- *σ restricted to every branch is an additive mapping, and*
- *for every nodes $x \sqsubset y$, $t \models \Phi(x,y)$ iff $\sigma(x,y) \in A_\Phi$.*

Furthermore, σ is monadically definable: for every $s \in S_\Phi$, there exists a monadic formula $\Phi_s(x,y)$ such that for every tree t and nodes $x \sqsubset y$, $t \models \Phi_s(x,y)$ iff $\sigma(x,y) = s$.

And the result is then the following.

Theorem 9. *For every monadic interpretation $\mathcal{I}_{MSO}$, there exists a monadic marking $\mathcal{M}_{MSO}$ and a first-order interpretation $\mathcal{I}_{FO}$ such that for every tree t, $\mathcal{I}_{MSO}(t) = \mathcal{I}_{FO}(\mathcal{M}_{MSO}(t))$.*

Proof. Wlog, we prove the result for an interpretation $\mathcal{I}_{MSO}$ with a single formula $\Phi(x_1,\dots,x_n)$. Using Lemma 25, we just have to show how to obtain an equivalent to a formula of the form $x \sqsubset y \wedge \Psi(x,y)$ as the combination of a monadic marking and a first-order formula. For this, we use Lemma 26 which tells us that the value of $\Psi(x,y)$ can be uncovered by projection of an additive labelling. And we use Theorem 8 for reducing the computation of the additive labelling to the combination of a monadic marking and a first-order formula.

Note that this argument heavily relies on the determinism of the construction of Theorem 8. Indeed, one has to mark every branch of a tree, *a priori* with a different marking. The determinism property allows to have a single marking for the whole tree. □

7.4 Consequences for infinite structures

The goal of this section is to show how the results given above, namely Theorem 9, have direct new consequences in the definition of some families of finitely presentable infinite structures. There is no real technical contribution in this section but rather a presentation of those consequences to the theory of infinite structures. Let us warn the reader that we do not intend to provide a survey of this area, since this would require much more space and would be out of topic. We rather directly concentrate on providing Theorems 11 and 12. Essentially, those results show that for the standard caracterisation of the families of prefix-recognizable graphs, as well as for the Caucal hierarchy, one can replace the monadic interpretations by first-order ones.

The prefix-recognizable graphs were introduced by Caucal via an internal definition [11]. Namely, fix a finite alphabet A. A *prefix-recognizable* graph is an infinite directed graph defined as follows. Its set of vertices is a regular language over the alphabet A. And each edge relation is a finite union of relations of the form $(U \times V).W$ with

$$(U \times V).W = \{(uw, vw) \,:\, u \in U,\, v \in V,\, w \in W\}\ ,$$

for U, V, W regular languages. By extension, a graph is *prefix recognizable* if it is isomorphic to such a graph. An important property of those graphs is that their monadic theory is decidable (this fact is due to Caucal [11]; it can be easily seen as a direct consequence of Rabin Theorem [21] stating that the complete binary tree has a decidable monadic theory, together with Theorem 10 below).

There exists different caracterisations for this class of graphs. We will use below the following one:

Theorem 10 (Blumensath [3]). *A graph is prefix-recognizable iff it is isomorphic to a monadic interpretation of the complete binary tree.*

Using this theorem as guide, one can extend the definition of prefix-recognisability to relational structures: we call a relational structure *prefix-recognizable* if it is monadically interpretable in the complete binary tree.

Theorem 9 provides another — new — caracterisation of prefix-recognizable structures, Theorem 11. Beforehand, we need the following lemma.

Lemma 27. *Let t be a regular tree. Then there exists a first-order interpretation $\mathcal{I}_{FO}$ such that t is isomorphic to $\mathcal{I}_{FO}(\Delta_2)$.*

Proof. It is sufficient to consider that the regular tree is the complete binary tree together with a regular labelling in some finite alphabet A attached to every node. This means that there exists a deterministic and complete finite automata $\mathcal{A}$ of finite words over the alphabet $\{0, 1\}$, with each state labelled by a letter in A, such that the label of a node u is the letter attached to the sole state reached from the initial state while reading u. Let this automaton have states Q, initial state q_0, and transition function δ from $Q \times \{0, 1\}$ to Q. As usual we extend this transition function into a mapping from $Q \times \{0, 1\}^*$ to Q. Wlog we can assume that there exists also a mapping d from Q to $\{0, 1\}$ such that for every state q in Q and letter a in $\{0, 1\}$, $d(\delta(q, a)) = a$; i.e. the automaton remembers whether the current node is a left or a right child.

Let n be a mapping numbering the states of $\mathcal{A}$ from 1 to $|Q|$. Given a word $u = a_1 a_2 \dots a_n$, the a_i's being letters in $\{0, 1\}$, define:

$$f(u) = 10^{n(q_0)}10^{n(q_1)}1 \dots 10^{n(q_n)}1$$

in which $q_0, q_1, \ldots, q_n$ are the $n+1$ states successively assumed by the automaton while reading the word u. The proviso concerning the mapping d makes f an injective mapping.

The image of f is first-order definable (as a language of words). Indeed, in order to verify that a word belongs to the image of f, it is sufficient to check a) that $10^{n(q_0)}1$ is a prefix, b) that the last letter is 1, and c) that every factor of the form 10^n10^m1 is such that $n = n(p)$ and $m = n(q)$ for some transition $\delta(p,a) = q$. Those verifications are first-order definable. Furthermore, for every word u, the state $\delta(q_0, u)$ is nothing but the sole state q such that $10^{n(q)}1$ is suffix of $f(u)$. This is also first-order definable.

From those remarks, it is easy to give a first-order interpretation which, given the complete binary tree, selects the nodes belonging to the image of f, and labels every node $f(u)$ by the state $\delta(q_0, u)$. This interpretation provides a new tree t'. Since all the relevant information — the label of the node, and its right-child/left-child nature — is encoded in each state, it is easy to first-order interpret t in t'. □

Theorem 11. *A structure is prefix-recognizable iff it is isomorphic to the first-order interpretation (with ancestor relation) of the complete binary tree.*

Proof. We have to show that given a monadic interpretation $\mathcal{I}_{MSO}$, there exists a first-order interpretation $\mathcal{I}_{FO}$ such that $\mathcal{I}_{MSO}(\Delta_2)$ is isomorphic to $\mathcal{I}_{FO}(\Delta_2)$. Using Theorem 9, we have that $\mathcal{I}_{MSO}(\Delta_2)$ is equal to $\mathcal{I}'_{FO}(\mathcal{L}_{MSO}(\Delta_2))$ for some monadic labelling $\mathcal{L}_{MSO}$ and first-order interpretation $\mathcal{I}'_{FO}$. Then using Lemma 27, we obtain an interpretation $\mathcal{I}''_{FO}$ such that $\mathcal{I}''_{FO}(\Delta_2)$ is isomorphic to $\mathcal{L}_{MSO}(\Delta_2)$. By closure of first-order interpretation under composition, $\mathcal{I}_{FO} = \mathcal{I}'_{FO} \circ \mathcal{I}''_{FO}$ is a first-order interpretation such that $\mathcal{I}_{FO}(\Delta_2)$ is isomorphic to $\mathcal{I}_{MSO}(\Delta_2)$. □

A similar approach can be used for caracterising the Caucal hierarchy. The Caucal hierarchy [12] is an extension of prefix-recognizable graphs to 'higher-order'. We use here the caracterisation of Carayol and Wöhrle [8] as a definition:

- The structures in Struct_0 are the finite relational structures.
- The graphs in Graph_n are the structures in Struct_n having a graph signature.
- The trees in Tree_{n+1} are the unfolding of graphs in Graph_n.
- The structures in Struct_{n+1} are the monadic interpretations of trees in Tree_{n+1}.

Since both the monadic interpretation and the unfolding preserve the decidability of the monadic theory, the trees, graphs and structures in the classes defined above have a decidable monadic theory.

The following interpretation shows that in the definition of this hierarchy, the monadic logic can be replaced by first-order logic.

Theorem 12. *The structures in Struct_n are, up to isomorphism, the first-order interpretation of trees in Tree_n.*

In fact, this is a direct consequence of Theorem 9 together with the following proposition (see [8], Proposition 1).

Proposition 2. *The class Tree_n is closed under monadic markings.*

Acknowledgement

I am deeply grateful to Achim Blumensath and Olivier Carton for their help in the production of this document.

References

1. J. R. Bchi. 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.
2. N. Bedon and O. Carton. An eilenberg theorem for words on countable ordinals. In *LATIN*, pages 53–64, 1998.
3. A. Blumensath. Prefix-recognisable graphs and monadic second-order logic. Technical Report AIB-06-2001, RWTH Aachen, May 2001.
4. A. Blumensath. A short combinatorial proof of Rabin's theorem. Personal communication, 2006.
5. Mikolaj Bojańczyk and Thomas Colcombet. Bounds in omega-regularity. In *IEEE Symposium on Logic In Computer Science*, pages 285–296, 2006.
6. T. C. Brown. An interesting combinatorial method in the theory of locally finite semigroups. *Pacific Journal of Mathematics*, 36(2):277–294, 1971.
7. V. Bruyre and O. Carton. Automata on linear orderings. In *MFCS*, volume 2136, pages 236–247, 2001.
8. Arnaud Carayol and Stefan Wöhrle. The Caucal hierarchy of infinite graphs in terms of logic and higher-order pushdown automata. In *FSTTCS'03*, volume 2914 of *LNCS*, pages 112–123. Springer, 2003.
9. O. Carton and C. Rispal. Complementation of rational sets on scattered linear orderings of finite rank. In *LATIN*, pages 292–301, 2004.
10. O. Carton and C. Rispal. Complementation of rational sets on countable scattered linear orderings. *Int. J. Found. Comput. Sci.*, 16(4):767–786, 2005.
11. Didier Caucal. On infinite transition graphs having a decidable monadic theory. In *ICALP'96*, volume 1099 of *LNCS*, pages 194–205. Springer, 1996.
12. Didier Caucal. On infinite terms having a decidable monadic theory. In *MFCS'02*, volume 2420 of *LNCS*, pages 165–176. Springer, 2002.
13. J. Chalopin and H. Leung. On factorization forests of finite height. *Theoretical Computer Science*, 310(1–3):489–499, jan 2004.
14. G. Lallement. *Semigroups and Combinatorial Applications*. Wiley, New-York, 1979.
15. R. McNaughton. Testing and generating infinite sequences by a finite automaton. *Information and Control*, 9(5):521–530, 1966.
16. D. Perrin and J.E. Pin. *Semigroups,Formal Languages and Groups*, chapter Semigroups and automata on infinite words, pages 49–72. Kluwer, 1995.
17. J-E. Pin and P. Weil. Polynominal closure and unambiguous product. *Theory Comput. Syst.*, 30(4):383–422, 1997.
18. J.E. Pin. *Variétés de langages formels*. Masson, Paris, 1984.
19. J.E. Pin. *Varieties of formal languages*. North Oxford, London and Plenum, New-York, 1986.
20. J.E. Pin, B. le Sac, and P. Weil. Semigroups with idempotent stabilizers and application to automata theory. *Int. J. of Alg. and Comput.*, 1(3):291—314, 1991.
21. M.O. Rabin. Decidability of second-order theories and automata on infinite trees. *Trans. Amer. Math. soc.*, 141:1–35, 1969.
22. J. G. Rosenstein. *Linear Orderings*. Academic Press, New York, 1982.
23. S. Shelah. The monadic theory of order. *Annals Math*, 102:379–419, 1975.
24. I. Simon. Factorization forests of finite height. *Theor. Comput. Sci.*, 72(1):65–94, 1990.
25. 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.