# A combinatorial theorem for trees
## Applications to monadic logic and infinite structures

Thomas Colcombet

Cnrs/Irisa
thomas.colcombet@irisa.fr

**Topics:** Semigroups, Ramseyan factorisation,
Monadic second-order logic, Trees, Infinite structures.

**Abstract.** Following the idea developed by I. Simon in his theorem of Ramseyan factorisation forests, we develop a result of 'deterministic factorisations'. This extra determinism property makes it usable on trees (finite or infinite).
We apply our result for proving that, *over trees*, every monadic interpretation is equivalent to the composition of a first-order interpretation (with access to the ancestor relation) and a monadic marking. Using this remark, we give new characterisations for prefix-recognisable structures and for the Caucal hierarchy.
Furthermore, we believe that this approach has other potential applications.

## 1 Introduction

The theorem of factorisation forests was proposed by Simon [20]. One way to present it is the following. For every semigroup morphism $\varphi$ 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 original statement is slightly different in its presentation. It establishes the existence of a so called 'Ramseyan factorisation of bounded depth' for every word; those factorisations intuitively witness the acceptance of the word by the regular expression mentioned above. The present paper is based on the proof of the theorem of factorisation forests in [9] which is a simplification of the original presentation.

The result itself has been used for various applications. In [21], Simon uses this theorem for studying the finiteness problem of regular languages of matrices over the tropical semiring (i.e. the semiring $\mathbb{N} \cup \{\infty\}$ equipped with the minimum and addition operations). This problem is equivalent to the limitedness problem for distance automata. This question is at the heart of the very difficult proof of decidability of the star-height problem due to Hashigushi [11] (the star-height problem consists in determining how many nesting of Kleene stars are required for describing a given regular language of words by a regular

expression). In [16] the theorem is used in a characterisation of the polynomial closure of a variety of languages. In [2], the authors use the theorem of factorisation forests in a complementation result extending the one of Büchi over infinite words. A direct consequence of the result in [2] is the decidability of the limitedness problem for nested distance desert automata: this problem extends the one for distance automata seen above, and is the cornerstone of the modern and much simpler solution to the star-height problem proposed by Kirsten [12]. In general the theorem of factorisation forests entails very deep consequences in the understanding of the structure of semigroups. For instance, one directly derives from it a constructive proof of Brown's lemma on locally finite semigroups [3].

Independently of the contributions of this paper, let us advertise the importance of the factorisation forests theorem. This result is well known in semigroup theory, in which some of its consequences are investigated. But this theorem clearly has other potential fields of application. In the present paper for instance, we use the approach for an application in logic. The interest of this theorem is in fact more natural outside the scope of semigroup theory: for a non-specialist in semigroups, it happens to be much easier to use than to prove. Thus, it is hardly avoidable in some situations (such as in [2]).

The present paper is an attempt to adapt the theorem of factorisation forests in a framework suitable for its use on trees. Essentially the problem we have concerning the original statement is the following: given two words sharing a common prefix, the factorisation forests theorem explicits the existence of a factorisation for each of the two words, but those two factorisations need not coincide on the common prefix. For eliminating this problem, we introduce an extra determinism requirement: the original theorem shows the existence of a factorisation for every word; our theorem shows the existence of a factorisation which is computable 'deterministically, on-line' while reading the word from left to right. For reaching this goal, we modify in two ways the original result. A cosmetic modification is that we drop the original formalism using trees — determinism does not fit naturally in it —, and replace it by the notion of splits for representing factorisation (see below). The second modification consists in weakening the hypothesis of 'Ramseyanity', and replace it by a notion of 'forward Ramseyanity'. Without this weakening, the result would simply not hold.

The second part of the paper is devoted to an application of this result to monadic (second-order) logic over trees. This result has been chosen as an application because it is it is new, because it does not contain too much technicalities, and also because it could not be derived from weaker version of the main theorem. Let us recall that the monadic logic is an extension of first-order logic by the possibility to quantify over sets of elements. Over words, trees as well as infinite words and trees (of length/height $\omega$), the expressivity of closed monadic formulæ coincide with the standard classes of automata ([5, 4, 17]). In particular, this logic is known to be more expressive than first-order logic, already over words. We use our result to decompose monadic formulæ over trees: every monadic formula with only free first-order variables is equivalent *over trees* to a first-order formula with access to the ancestor relation and to monadically

defined unary predicates. Equivalently, every monadic interpretation is equivalent, over trees, to the composition of a first-order interpretation and a monadic marking.

We apply this result to the theory of infinite structures. We give new characterisations to the class of prefix-recognisable structures as well as to the Caucal hierarchy.

## 2 Main result

We first define semigroups and additive labellings, then words in Sections 2.1 and 2.2. Our main result is presented in Section 2.3.

### 2.1 Semigroups and Additive Labellings

A *semigroup* $(S, .)$ is a set $S$ equipped with an associative binary operator written multiplicatively. Groups and monoids are particular instances of semigroups. A *morphism of semigroup* from a semigroup $(S, .)$ to a semigroup $(S', .')$ is a mapping $\varphi$ 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 us recall that a *linear ordering* $(\alpha, <)$ is a set $\alpha$ together with a total strict ordering relation $<$. Let $\alpha$ be a linear ordering and $(S, .)$ be a semigroup. A mapping $\sigma$ 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 $\alpha$, $\sigma(x, y).\sigma(y, z) = \sigma(x, z)$.

### 2.2 Words

Given an alphabet $A$, we denote by $A^*$ the set of finite words over $A$, i.e. finite sequences of letters in $A$. The length of the word is the length of the sequence. The empty word is $\varepsilon$, and $A^+$ represents $A^* \setminus \{\varepsilon\}$. $A^+$ equipped with the concatenation of words is a semigroup. Given a word $u$ of length $n$, and $i, j$ with $0 \leq i \leq j \leq n$, $u_{i,j}$ is the word $u_{i+1}u_{i+2}\dots u_j$. Given a finite semigroup $S$, a morphism of semigroup $\varphi$ from $A^+$ to $S$, and a word $u$ of length $n$, $\varphi_u$ is the *additive labelling* from $[0, n]$ to $S$ defined by

$$\varphi_u(i, j) = \varphi(u_{i,j}).$$

Reciprocally, given an additive labelling $\sigma$ over $([0, n], <)$ for some $n$, one can associate the word $\langle\sigma\rangle$ of length $n$ over the alphabet $S$, the $i$th letter of which is $\sigma(i-1, i)$. Of course, $\varphi_{\langle\sigma\rangle} = \sigma$, where $\varphi$ is the canonical semigroup morphism from $S^+$ to $S$. According to this remark, additive labellings and words together with a semigroup morphism form two sides of the same object.

### 2.3 Main Theorem

A *split of height* $N$ of a linear ordering $\alpha$ is a mapping $s$ from $\alpha$ to $[1, N]$ (we use square brackets for intervals of natural numbers). Given a split, two elements $x$ and $y$ in $\alpha$ 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 $s^{-1}(k)$. A split $s$ of height $N$ is *forward Ramseyan* wrt. $\sigma$ if for every $k = 1 \dots n$ and every $x, y, x', y'$ in the same class of $k$-neighbourhood with $x < y$ and $x' < y'$,

$$\sigma(x, y) = \sigma(x, y).\sigma(x', y') \ . \tag{1}$$

So in particular, $\sigma(x, y)$ is an idempotent, but $\sigma(x, y)$ and $\sigma(x', y')$ may be different idempotents[1]. Our main result is the following.

**Theorem 1.** *Fix a finite semigroup $(S, .)$, an alphabet $A$ and a semigroup morphism $\varphi$ from $A^+$ to $S$. There is a partition of $A^*$ into regular languages $L_1, \dots, L_K$ with $K \leq |S|$ such that for every word $u$ of length $n$, $s_u$ defined by*

$$\text{for all } i \in [0, n], \quad s_u(i) = k \quad \text{such that } u_{0,i} \in L_k,$$

*is forward Ramseyan for $\varphi_u$.*

The proof essentially follows the one of Chalopin and Leung [9]. In particular, it is based on a decomposition of the semigroup following Green's relations. Green's relation reflects the interplay of ideals in a semigroup [10], and their use provides deep informations on the structure of the semigroup.

*Example 1.* For simplicity, we identify $A$ with $S$, and $\varphi$ is the identity over letters. Let $S$ be $\{a, b, c\}$ together with the product defined by $a = ab = aa$, $b = ba = bb$ and $c = cc = ac = bc = ca = cb$, then the languages

$$L_1 = \varepsilon + (a + b)^* c \ , \qquad \text{and} \quad L_2 = (a + b + c)^*(a + b)$$

form a valid output of the theorem. For instance, consider the word *abcbaabacbaa*, the split defined is (the value of the split being interleaved in the word):

$$1\, a\, 2\, b\, 2\, c\, 1\, b\, 2\, a\, 2\, a\, 2\, b\, 2\, a\, 2\, c\, 1\, b\, 2\, a\, 2\, a\, 2 \ .$$

Given two 1-neighbour positions $i < j$, the letter just before $j$ is $c$; this means $\varphi_u(i, j) = c$. While given two 2-neighbour positions $i < j$, all letters in $u_{i,j}$ belong to $\{a, b\}$; this means $\varphi_u(i, j) \in \{a, b\}$. Those remarks entail the forward Ramseyanity since $c$ is an idempotent, and $a, b$ are idempotents satisfying $ab = a$ and $ba = b$.

## 3 Application to Monadic Second-Order Logic

We first define structures, graphs and trees in Section 3.1. Then Section 3.2 introduces logics. Section 3.3 presents our result, Theorem 2, and Section 3.4 studies its consequences over infinite structures.

[1] In terms of Green's relation, $\sigma(x, y)$ and $\sigma(x', y')$ are $\mathcal{L}$-equivalent idempotents.

### 3.1 Structures

**Structures.** A *(relational) structure* $(\mathcal{U}, R_1, \ldots, R_n)$ is a set $\mathcal{U}$, called the *universe*, together with *relations* $R_1, \ldots, R_n$ of fixed finite arity over $\mathcal{U}$. Each relation $R$ has a *name* that we write $R$ itself. The relation is called the *interpretation* of the name in the structure. The *signature* of a structure contains the names involved together with their arity. By extension, we allow (partial) functions from $\mathcal{U}^k$ to some finite set $E = \{e_1, \ldots, e_n\}$. Such a function $f$ is nothing but a shorthand for using $n$ $k$-ary relations $F_1, \ldots, F_n$, each $F_i$ being interpreted as $f^{-1}(e_i)$.

**Words.** Our base structure is $([0, n], <)$, i.e. the natural numbers equipped with the natural ordering. An additive labelling $\sigma$ on it to some finite semigroup can be directly represented in it according to the remark above: this provides the structure $([0, n], <, \sigma)$. Our coding of a word $u$ of length $n$ is slightly non-standard. It is the structure $([0, n], <, u)$ where $u$ is a partial mapping from $[1, n]$ to $A$. The element 0 of this structure is unused. This makes it easier to jump from $([0, n], <, \sigma)$ to $([0, n], <, \langle\sigma\rangle)$ and vice versa.

**Graphs.** A *(directed) graph* is a structure for which all relations have arity 1 but one of arity 2 called the *edge relation*. The elements of the universe are called *vertices*, the unary relations are *labelling relations*. 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*.

**Trees.** A *tree* $t$ is a graph for which the edge relation is called the *ancestor relation*, is denoted by $\sqsubseteq$, and satisfies:

- 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 finite and totally ordered.

The vertices of a tree are called *nodes*. Maximal chains are called *branches*. The maximal element smaller than node $u$ is called (when it exists) the *parent* of $u$.

We extend the notions for words to trees: an *additive labelling* is a mapping $\sigma$ from pairs of nodes $(x, y)$ such that $x \sqsubset y$ to a finite semigroup $S$ which is an additive labelling when restricted to every branch. We also note by $\langle\sigma\rangle$ the partial function from nodes different from the root to $S$ defined by $\langle\sigma\rangle(u) = \sigma(v, u)$ for $v$ the parent of $u$ (if it exists). A *split* of height $N$ is a mapping from nodes to $[1, N]$. The split is *forward Ramseyan* wrt. $\sigma$ if it is forward Ramseyan wrt. $\sigma$ over every branch.

*Caution:* The trees are *not* defined by a 'direct successor' relation, but rather by the ancestor relation. This has major impact on the logic 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 ancestor relation is necessary in this work.

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 $\Delta_2$ 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 paths of origin $v$, as ancestor relation the prefix relation over paths, and such that a path $\pi$ is labelled by $a$ in the unfolding iff its destination is labelled by $a$ in the graph.

### 3.2 Logics

*First-order logic.* We assume a countable set of *first-order variables* $x, y, \ldots$ to pick from. The *atomic formulæ* are $R(x_1, \ldots, x_n)$ for $x_1, \ldots, x_n$ first-order variables and $R$ a 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 atomic formulæ combined by the boolean connectives $\vee, \wedge, \neg$, and the first-order quantifiers $\exists x$ and $\forall x$.

*Monadic logic.* We assume 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 first-order and $X$ monadic.

We use the standard notion of *free variables*. A formula without free variables is *closed*. We denote by $\mathcal{S} \models \phi$ the fact, for a closed formula $\phi$ and a structure $\mathcal{S}$, that the formula is true over the structure $\mathcal{S}$. The value of first-order variables range over elements of the universe of the structure, while monadic variables take as values subsets of the universe. For $\mathcal{S} \models \phi$, we also say that $\mathcal{S}$ is a *model* of $\phi$, or that $\phi$ is *satisfied* over $\mathcal{S}$. We also allow ourselves to write $\phi(x_1, \ldots, x_n)$ to denote that the free-variables of $\phi$ 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 $\phi$ is true over the structure $\mathcal{S}$, using the valuation mapping $x_i$ to $u_i$.

A 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 $\phi$ of the logic $L$, answers whether $\mathcal{S} \models \phi$ or not.

**Definability.** Let $R$ be a relation of arity $k$ over the universe of some structure $\mathcal{S}$. It is said $L$ *definable* (where $L$ is either first-order or monadic) if there exists an $L$ formula $\phi(x_1, \ldots, x_k)$ such that $R(u_1, \ldots, u_k)$ iff $\mathcal{S} \models \phi(u_1, \ldots, u_k)$. *Implicitly*, when we refer to definability, we mean that the formula does not depend of the structure.

A structure $\mathcal{S}'$ is $L$ *definable* in $\mathcal{S}$ if its universe is an $L$ definable subset of the universe of $\mathcal{S}$, and all the relations in $\mathcal{S}'$ are $L$ definable in $\mathcal{S}$. We write $\mathcal{S}' \leq_{FO} \mathcal{S}$ (*resp.* $\mathcal{S}' \leq_{MSO} \mathcal{S}$) if $\mathcal{S}'$ if first-order (*resp.* monadically) definable in

$\mathcal{S}$. The special case $\mathcal{S}' \leq^1_{MSO} \mathcal{S}$ signifies that $\mathcal{S}'$ is the structure $\mathcal{S}$ augmented with some new monadically definable unary relations. We also say that $\mathcal{S}'$ is obtained by a *monadic marking* from $\mathcal{S}$. The relations $\lesssim_{FO}$, $\lesssim_{MSO}$ and $\lesssim^1_{MSO}$ correspond to $\leq_{FO}$, $\leq_{MSO}$ and $\leq^1_{MSO}$ respectively, up to isomorphism.

### 3.3 Main result

We prove in this section Theorem 2: every monadic interpretation is the composition of a first-order interpretation with a monadic marking. For simplicity, we use the notations $\leq_{FO}, \leq_{MSO}$, etc. But let us stress that the constructions are uniform in the sense that implicitly the formulæ involved do not depend on the structures, but only on their signatures.

We need two intermediate lemmas. Both can be proved using elementary compositional methods (see e.g., [19]). The first one establishes that a monadic formula with many first-order variables can be "first-order" reconstructed out of monadic formulæ of two free first-order variables.

**Lemma 1.** *Every monadic formula $\Phi(x_1, \ldots, x_n)$ is equivalent* over trees *to a first-order formula with access to binary relations defined by monadic formulæ of the form $x \sqsubset y \wedge \Psi(x, y)$.*

The second lemma states that a monadic formula of two free first-order variables can be seen as a monadically definable additive labelling.

**Lemma 2.** *For every monadic formula of the form $\Phi(x, y)$ of free first-order variables $x, y$, there exists a finite semigroup $S_\Phi$, a subset $A_\Phi \subseteq S_\Phi$, and an additive labelling $\sigma$ monadically definable in every tree $t$ such that*

$$\text{for every nodes } u \sqsubset v \text{ in } t, \quad t \models \Phi(u, v) \text{ iff } \sigma(u, v) \in A_\Phi \ .$$

The combination of the two previous lemmas yields the following.

**Corollary 1.** *For every relation $R$ monadically definable in some tree $t$,*

$$(t, R) \leq_{FO} (t, \sigma) \leq_{MSO} t \ ,$$

*where $\sigma$ is an additive labelling from $t$ to some finite semigroup.*

*Proof.* (Idea) If $R$ is defined by a monadic formula of the form $x \sqsubset y \wedge \Psi(x, y)$, this is a direct application of Lemma 2. Else, we decompose the formula defining $R$ as in Lemma 1, and use the argument above for coding each formula of the form $x \sqsubset y \wedge \Psi(x, y)$. Each time we obtain a semigroup and an additive labelling. By product, we combine all those informations into a single semigroup and a single additive labelling $\sigma$. □

The following lemma is the key argument. It shows how to first-order reconstruct an additive labelling $\sigma$ out of its unary presentation $\langle\sigma\rangle$, providing a forward Ramseyan split is given.

**Lemma 3.** *Fix a semigroup $S$. Then*

$$([0,n],<,\sigma) \leq_{FO} ([0,n],<,\langle\sigma\rangle,s)$$

*where $n$ is some natural number, $\sigma$ an additive labelling from $[0,n]$ to $S$, and $s$ a split of $[0,n]$ of height at most $|S|$ forward Ramseyan wrt. $\sigma$.*

*Proof.* We have to first-order define $\sigma$ in $([0,n],<,\langle\sigma\rangle,s)$. By a downward induction on $k = |S|+1,\ldots,1$, we construct a function $a_k(i,j)$ to $S$ such that for every $i<j$ in $[0,n]$, $a_k(i,j) = \sigma(i,j)$ whenever $s(x) \geq k$ for all $x \in ]i,j[$.

For $k = |S|+1$, then $j = i+1$. And $a_{|S|+1}(i,j) = \langle\sigma\rangle(j) = \sigma(i,j)$.

For $k \leq |S|$. Let $i<j$ lie in $[0,n]$, define

$$a_k(i,j) = \begin{cases} a_{k+1}(i,j) & \text{if } s(x) > k \text{ for all } x \in ]i,j[\ , \\ a_{k+1}(i,x).a_{k+1}(x,j) & \text{if } x \text{ is the only element in } ]i,j[ \\ & \text{such that } s(x) = k, \\ a_{k+1}(i,x).a_{k+1}(x,y).a_{k+1}(z,j) & \\ & \text{for } x<y\leq z \text{ in } ]i,j[ \\ & \text{with } s(x)=s(y)=s(z)=k \\ & \text{and } s(w)>k \text{ for all } w \in ]i,x[\cup]x,y[\cup]z,j[\,. \end{cases}$$

Let $i<j$ be in $[0,n]$ such that $s(x) \geq k$ for all $x \in ]i,j[$. We have to show $a_k(i,j) = \sigma(i,j)$. In the two first cases, the correctness is obvious. In the last case, $x,y,z$ are $k$-neighbours. Hence, by forward Ramseyanity of $s$, $\sigma(x,z) = \sigma(x,y).\sigma(y,z) = \sigma(x,z)$ (this holds even if $y=z$). Hence

$$\begin{aligned} a_k(i,j) &= a_{k+1}(i,x).a_{k+1}(x,y).a_{k+1}(z,j) \\ &= \sigma(i,x).\sigma(x,y).\sigma(z,j) \\ &= \sigma(i,x).\sigma(x,z).\sigma(z,j) \\ &= \sigma(i,j)\ . \end{aligned}$$

Those constructions are first-order definable. The result follows. □

The determinism allows to transfer easily this result to trees.

**Corollary 2.** *Fix a semigroup $S$. Then*

$$(t,\sigma) \leq_{FO} (t,\langle\sigma\rangle,s)$$

*for a tree $t$, an additive labelling $\sigma$ from $t$ to $s$, and a split of $t$ of height $|S|$ forward Ramseyan for $\sigma$.*

*Proof.* (Idea) The formula defining $\sigma(u,v)$ for $u \sqsubset v$ is obtained by relativisation of the formula obtained by Lemma 3 to $\{w \ : \ w \sqsubseteq v\}$. □

**Lemma 4.** *Let $R$ be a relation monadically definable in a tree $t$,*

$$(t,R) \leq_{FO} t' \leq^1_{MSO} t \quad \textit{for some tree } t'.$$

*Proof.* From Corollary 1 it is sufficient to derive from $(t,\sigma) \leq_{MSO} t$ that

$$(t,\sigma) \leq_{FO} t' \leq^1_{MSO} t$$

where $\sigma$ is an additive labelling to a finite semigroup $S$. Let $L_1, \ldots, L_{|S|}$ be as in Theorem 1. Set $s$ to be the split of $t$ defined by $s(u) = n$ such that $t|_{\{v \,:\, v \sqsubseteq u\}} \in L_n$ (we see here $t|_{\{v \,:\, v \sqsubseteq u\}}$ as a word, up to isomorphism). By Theorem 1, $s$ is forward Ramseyan wrt. $\sigma$. Furthermore, from the equivalence between regular languages and monadic logic, $s$ is monadically definable in $(t,\sigma)$. Obviously, $\langle\sigma\rangle$ is also monadically definable in $(t,\sigma)$. Combined with Corollary 2 we obtain:

$$(t,\sigma) \leq_{FO} (t, \langle\sigma\rangle, s) \leq_{MSO} (t,\sigma) \leq_{MSO} t \ .$$

Hence,

$$(t,\sigma) \leq_{FO} (t, \langle\sigma\rangle, s) \leq^1_{MSO} t \ .$$

□

By extension of Lemma 4 to structures, we obtain the expected theorem.

**Theorem 2.** *If $\mathcal{S} \leq_{MSO} t$ then $\mathcal{S} \leq_{FO} t' \leq^1_{MSO} t$ for some tree $t'$.*

### 3.4 Consequences for infinite structures

The goal of this section is to show how Theorem 2 has direct new consequences in the definition of some families of finitely presentable infinite structures: Theorems 4 and 5. But we do not intend to survey this area.

Prefix-recognisable graphs were introduced in [7]. Fix a finite alphabet $A$. A *prefix-recognisable* graph is an (possibly infinite) 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-recognisable* 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 [7]; it can be easily seen as a direct consequence of Rabin Theorem [17] stating that the complete binary tree has a decidable monadic theory, together with Theorem 3 below).

There exists different characterisations for this class of graphs. We will use below the following one due to Blumensath [1]:

**Theorem 3.** *A graph $G$ is prefix-recognisable iff $G \lesssim_{MSO} \Delta_2$.*

Following this idea, one extends prefix-recognisability to structures: a structure $\mathcal{S}$ is *prefix-recognisable* if $\mathcal{S} \lesssim_{MSO} \Delta_2$.

Theorem 4 provides another – new – characterisation to the prefix-recognisable structures. Beforehand, we need Lemma 5 stating that every regular binary tree is first-order definable in $\Delta_2$.

**Lemma 5.** *If* $t \leq^1_{MSO} \Delta_2$ *then* $t \lesssim_{FO} \Delta_2$.

And we obtain.

**Theorem 4.** *A structure* $\mathcal{S}$ *is prefix-recognisable iff* $\mathcal{S} \lesssim_{FO} \Delta_2$.

*Proof.* From $\mathcal{S} \lesssim_{MSO} \Delta_2$ and Theorem 2, $\mathcal{S} \lesssim_{FO} t' \leq^1_{MSO} \Delta_2$ for some $t'$. By Lemma 5, $\mathcal{S} \lesssim_{FO} t' \lesssim_{FO} \Delta_2$. Hence $\mathcal{S} \lesssim_{FO} \Delta_2$. The converse is obvious. □

A similar approach can be used for characterising the Caucal hierarchy [8], i.e a form of extension of prefix-recognisable graphs to 'higher-order'. We use here the characterisation in [6] as a definition:

- The structures in $Struct_0$ are the finite structures.
- The graphs in $Graph_n$ are the graphs[2] in $Struct_n$.
- The trees in $Tree_{n+1}$ are the unfolding of graphs in $Graph_n$.
- A structure $\mathcal{S}$ is in $Struct_{n+1}$ if $\mathcal{S} \lesssim_{MSO} t$ for some tree $t$ in $Tree_{n+1}$.

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

**Theorem 5.** $\mathcal{S}$ *is in* $Struct_{n+1}$ *iff* $\mathcal{S} \lesssim_{FO} t$ *for some tree* $t$ *in* $Tree_{n+1}$.

In fact, this is a direct combination of the definitions, of Theorem 2, and of the following proposition (see [6], Proposition 1).

**Proposition 1.** *For* $t$ *in* $Tree_n$, *every* $t' \lesssim^1_{MSO} t$ *is also in* $Tree_n$.

## 4 Other Consequences, Perspectives

### Determinisation of Regular Languages of Infinite words

From our result can also be derived McNaughton's determinisation theorem [13]. This theorem states that for every regular language of infinite words of length $\omega$ (regular meaning accepted by a Büchi automaton), there exists a *deterministic parity automaton* which accepts the same language. Such an automaton is a standard finite deterministic automaton without final states, the states of which are labelled by natural numbers among $1, \ldots, p$ called their *priorities*. An infinite word is accepted by such an automaton if the least priority appearing infinitely often in the (unique by determinism) run is even. The deteterminisation result is fundamental in the theory of languages of infinite words. In particular, it is used in most proofs of the theorem of Rabin.

Let us sketch the link with our result. Given a Büchi automaton accepting a language $L$, it is natural to associate to it a structure of finite semigroup $S$ as well as a morphism $\varphi$ from finite words to $S$ such that it is sufficient to know $\varphi(u_1), \varphi(u_2), \ldots$ for determining whether the word $u_1 u_2 \ldots$ belongs to $L$

[2] In the original version, graphs have their edges labelled. We drop this since it has no impact on the definition of $Struct_n$.

(this was already the approach of Büchi in his proof of complementation [4], see also [15] for the more explicit presentation via $\omega$-semigroups). In particular, given two words $u, v$, the membership of $uv^\omega = uvvv\cdots$ in $L$ does only depend of $\varphi(u)$ and $\varphi(v)$.

One can apply Theorem 1 to this semigroup, and obtain for each word $u$ the forward Ramseyan split $s_u$. One constructs a deterministic parity automaton which, when reading a word of $u$ of length $n$, reaches a state of priority:

$$\begin{cases} 2s_u(n) & \text{if } u_{0,m}u_{m,n}^\omega \in L \text{ for } m = \max\{m < n : s_u(n) = s_u(m)\}, \\ 2s_u(n)+1 & \text{if } u_{0,m}u_{m,n}^\omega \notin L, \text{ or } m \text{ does not exist.} \end{cases}$$

One checks that a) this can be implemented by a finite deterministic parity automaton, and b) that the language it accepts is $L$ (using forward Ramseyanity).

This construction yields a doubly exponential automaton in the size of the original automaton. It is comparable in complexity to the original proof of McNaughton, but quiet far from Safra's construction [18]. A clother study of the construction above shows that in fact only a third of the proof of Theorem 1 (the proof treats separately three different cases) is sufficient for establishing McNaughton's determinisation result. This means that determinisation does not illustrate the full interest of Theorem 1 and should be considered more as a side-effect.

Another combinatorial approach to determinisation of Büchi automata is known from [22]. The combinatorial lemma it involves is suitable for determinisation, but is not as expressive as Theorem 1. And in particular it is not sufficient for proving Theorem 2.

### Infinite trees

Another motivation for factorising trees is the perspective to give a new proof to the theorem of Rabin of decidability of monadic second-order logic over infinite trees [17]. Such a contribution would be an answer to a long lasting open question of Shelah [19]. It would also generalise the original proof of Büchi for infinite words to the case of infinite trees. Theorem 1 is far from being sufficient for such an application. A reason for that is the irreducibility of nondeterminism in tree automata (it is for instance shown in [14] that some languages of infinite trees are not accepted by any unambiguous automaton, i.e., an automaton having a single accepting run for each accepted input tree). Standard proofs of the result of Rabin handle this problem using game theory, see e.g., [23]. The theorem developed in this paper has no nondeterminism feature, and for this reason cannot be sufficient alone for this application.

Despite this, the main reason for the introduction of Theorem 1 by the author is its perspective of usage in an extension of the work in [2] to infinite trees, which would be at the same time an extension of Rabin's theorem

### Acknowledgement

Many thanks to Achim Blumensath who follows this work from the begining.

## References


1. A. Blumensath. Prefix-recognisable graphs and monadic second-order logic. Technical Report AIB-06-2001, RWTH Aachen, May 2001.
2. M. Bojańczyk and T. Colcombet. Bounds in omega-regularity. In *IEEE Symposium on Logic In Computer Science*, pages 285–296, 2006.
3. T. C. Brown. An interesting combinatorial method in the theory of locally finite semigroups. *Pacific Journal of Mathematics*, 36(2):277–294, 1971.
4. J.R. Büchi. On a decision method in restricted second order arithmetic. In *Proceedings of the International Congress on Logic, Methodology and Philosophy of Science*, pages 1–11. Stanford University press, 1960.
5. J.R. Büchi and C.C. Elgot. Decision problems of weak second order arithmetics and finite automata. *Notices of the American Math. Soc.*, 5:834, 1958.
6. A. Carayol and S. 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.
7. D. Caucal. On infinite transition graphs having a decidable monadic theory. In *ICALP'96*, volume 1099 of *LNCS*, pages 194–205. Springer, 1996.
8. D. Caucal. On infinite terms having a decidable monadic theory. In *MFCS'02*, volume 2420 of *LNCS*, pages 165–176. Springer, 2002.
9. J. Chalopin and H. Leung. On factorization forests of finite height. *Theoretical Computer Science*, 310(1–3):489–499, jan 2004.
10. J. A. Green. On the structure of semigroups. *Annals of Mathematics*, 54(1):163–172, 1951.
11. K. Hashiguchi. Relative star height, star height and finite automata with distance functions. In *Formal Properties of Finite Automata and Applications*, pages 74–88, 1988.
12. D. Kirsten. Distance desert automata and the star height problem. *RAIRO*, 3(39):455–509, 2005.
13. R. McNaughton. Testing and generating infinite sequences by a finite automaton. *Information and Control*, 9(5):521–530, 1966.
14. D. Niwinski and I. Walukiewicz. Ambiguity problem for automata on infinite trees. Personal communication, 1997.
15. D. Perrin and J-E. Pin. *Semigroups,Formal Languages and Groups*, chapter Semigroups and automata on infinite words, pages 49–72. Kluwer, 1995.
16. J-E. Pin and P. Weil. Polynomial closure and unambiguous product. *Theory Comput. Syst.*, 30(4):383–422, 1997.
17. M.O. Rabin. Decidability of second-order theories and automata on infinite trees. *Trans. Amer. Math. soc.*, 141:1–35, 1969.
18. S. Safra. On the complexity of omega-automata. In *FOCS*, pages 319–327, 1988.
19. S. Shelah. The monadic theory of order. *Annals Math*, 102:379–419, 1975.
20. I. Simon. Factorization forests of finite height. *Theor. Comput. Sci.*, 72(1):65–94, 1990.
21. I. Simon. On semigroups of matrices over the tropical semiring. *ITA*, 28(3-4):277–294, 1994.
22. W. Thomas. A combinatorial approach to the theory of $\omega$-automata. *Information and Control*, 48:261–283, 1981.
23. 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.