Factorisation Forests for Infinite Words
Application to Countable Scattered Linear Orderings

Thomas Colcombet

Cnrs/Irisa
thomas.colcombet@irisa.fr

Abstract. The theorem of factorisation forests shows the existence of nested factorisations — a la Ramsey — for finite words. This theorem has important applications in semigroup theory, and beyond.

We provide two improvements to the standard result. First we improve on all previously known bounds for the standard theorem. Second, we extend it to every ‘complete linear ordering’. We use this variant in a simplified proof of complementation of automata over words of countable scattered domain.

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

1 Introduction

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

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

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

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

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

2 Definitions

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

2.1 Linear Orderings

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

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