L3 Summer internship report
Laboratoire Bordelais de Recherche en Informatique
June – July 2009

Advisors: Anca Muscholl and Marc Zeitoun

Partial commutation closures of recognizable languages

by

Antoine Delignat-Lavaud
antoine@delignat-lavaud.fr

Contents

1 Acknowledgements and internship description

During my 7-weeks internship, I worked with the members of the *modeling and verification* workgroup in the *formal methods* team at the *Laboratoire Bordelais de Recherche en Informatique*.

I was given very good working conditions, having a lot of office space, a huge whiteboard and some air conditioning to make the customary summer heat of Bordeaux more bearable.

But, above all, it is the dedication of my advisors, Anca Muscholl and Marc Zeitoun, that made this internship an enlightning experience. I can not thank them enough for their guidance and willingness to involve me in their research.

This internship allowed me to familiarize with the very broad variety of tools and techniques, both from mathematics and computer science, that are developed in the mathematical theory of automata and formal languages.

2 Introduction to trace monoids and recognizability

In this section, we introduce the theory of traces and recognizability in non-free monoids. This has been an active field of research since the seventies, when it was popularized by Mazurkiewicz as a model of concurrency closely related to the well-developed theory of automata. This section was mainly inspired by the introduction in [1] by Diekert and Métivier.

2.1 Definition and basic properties

An alphabet Σ is a finite set whose elements are called letters. The free monoid over Σ is the set of all finite words over Σ, or equivalently the Kleene-star $\Sigma^* = \bigcup_{i \geqslant 0} \Sigma^i$; it is a monoid for concatenation where the empty word, which we denote 1, is the neutral element. Given a word $w \in \Sigma^*$, $u = a_1 \cdots a_n$ with $a_i \in \Sigma$ is a *factor* of w if there exist some $x, y \in \Sigma^*$ such that $w = xuy$. It is a *subword* of w if there exist some $w_0, \ldots, w_n \in \Sigma^*$ such that $w = w_0 u_1 w_1 \cdots u_n w_n$. If $w, u \in \Sigma^*$, $|w|$ denotes the length of w, $|w|_u$ the number of occurences of u as a factor in w and $\binom{w}{u}$ the number of occurences of u as a subword of w.

An *independence* or *commutation* relation is a symmetric and irreflexive relation $I \subseteq \Sigma \times \Sigma$. The pair (Σ, I) is called the *independence alphabet* and can be represented as an undirected graph called the *commutation graph*. The complement of I in $\Sigma \times \Sigma$ is called the dependence relation and is denoted D.

Example 1 $\Sigma = \{a, b, c, d\}$, $I = \{(a, b), (b, c), (c, d)\}$. *The commutation and dependence graphs are given below. They will be studied in depth in section 5.*

Given a commutation relation I, we define the *commutation equivalence* $\sim_I$ over Σ^* as the least congruence such that $ab \sim_I ba$ for all $(a,b) \in I$. The quotient $\mathbb{M}(\Sigma, I) = \Sigma^*/\sim_I$ (or $\mathbb{M}$ if there is no ambiguity) is called the *free partially I-commutative monoid* or *trace monoid*, elements of $\mathbb{M}$ are called *traces*. The canonical quotient homomorphism is denoted φ_I or $[\cdot]_I$. If $t \in \mathbb{M}$, a word $w \in \Sigma^*$ such that $t = [w]_I$ is called a *linearization* of t.

If for all $a \neq b \in \Sigma$, we have $(a,b) \in I$, $\mathbb{M}$ is called the *free commutative monoid*, and is isomorphic to $\mathbb{N}^\Sigma$ through the *Parikh homomorphism*:

$$\begin{array}{rccc} \pi : & \mathbb{M} & \longrightarrow & \mathbb{N}^\Sigma \\ & t = [w]_I & \longmapsto & (|w|_a)_{a\in\Sigma} \end{array}$$

In this case, the projection φ_I is simply denoted φ or $[\cdot]$.

A trace language is a subset $T \subseteq \mathbb{M}$. Note that it can also be viewed as a word language over Σ, namely $\varphi^{-1}(T)$. We say that a language $L \subseteq \Sigma^*$ is closed under I-commutation if $\varphi^{-1}([L]_I) = L$.

2.2 Recognizable trace languages

Definition 1 *Given a monoid M, an M-*automaton *is a tuple $\mathcal{A} = (Q, \delta, F)$ where Q is a finite monoid, $F \subseteq Q$ and δ is a monoid homomorphism from M to Q. The subset of M* recognized *by $\mathcal{A}$ is $\delta^{-1}(F)$. The family of subsets of M recognized by a M-automaton is denoted* $\mathrm{Rec}(M)$.

Definition 2 *Let M be a monoid and let $T \subseteq M$. The* syntactic preorder *over T, $\leqslant_T$ is defined by $x \leqslant_T y$ if for all $u, v \in M, uyv \in T \Rightarrow uxv \in T$. The syntactic equivalcne $\equiv_T$ is defined by $x \equiv_T y \Leftrightarrow x \leqslant_T y \wedge y \leqslant_T x$. The syntactic monoid of T is the quotient $M_T = M/\equiv_T$.*

Definition 3 *The family of rational sets of a monoid M, denoted* $\mathrm{Rat}(M)$*, is the closure under union, concatenation and iteration (Kleene star) of the set of finite subsets of M. The family of star-free sets, denoted $SF(M)$, is obtained by replacing iteration by complementation in the above definition.*

Proposition 1 *[2] Let $\varphi : \Sigma^* \to M$ be a surjective homomorphism and $T \subseteq M$. The following assertions are equivalent:*

1. $T \in \mathrm{Rec}(M)$
2. *$\equiv_T$ is of finite index, i.e. M_T is finite.*

3. $\varphi^{-1}(T)$ is a rational language over Σ.

Proof: (1) $\Rightarrow$ (2): Let (Q, δ, F) be a M-automaton that recognizes T. We denote by $\bar{x}$ the equivalence class of $x \in M$ for $\equiv_T$. Notice that $|\{\bar{x} \,|\, x \in M\}| \leqslant |Q|$: if $x, y \in M$ are such that $\delta(x) = \delta(y)$, then $\bar{x} = \bar{y}$ since:

$$\forall u, v \in M, uxv \in T \Leftrightarrow \delta(u)\delta(x)\delta(v) \in F \Leftrightarrow \delta(u)\delta(y)\delta(v) \in F \Leftrightarrow uyv \in T$$

(2) $\Rightarrow$ (1): If $\equiv_T$ is of finite index, let $\psi : M \to M_T$ be the canonical projection on the syntactic monoid, we define the M-automaton $\mathcal{A} = (M_T, \psi, \psi(T))$. The language recognized by $\mathcal{A}$ is $\psi^{-1} \circ \psi(T) \supseteq T$. But since T is saturated by $\equiv_T$, i.e. $\psi^{-1} \circ \psi(T) \subseteq T$, $\mathcal{A}$ recognizes T.
(1) $\Leftrightarrow$ (3): (Q, δ, F) is a M-automaton that recognizes T if and only if $(Q, \delta \circ \varphi, F)$ is a Σ^*-automaton that recognizes $\varphi^{-1}(T)$. $\diamondsuit$

Definition 4 *We say that a trace (or word) is connected if the set of letters it contains induces a connected subgraph of (Σ, D).*

Definition 5 *Assume the alphabet Σ is totally ordered by $<$. The* lexicographic normal form *of a trace, denoted $LNF(t)$, is the least linearization of t with regard to $<$.*

Proposition 2 *A word w is the lexicographic normal form of a trace $t \in \mathbb{M}$ if for all factorizations $w = xbyaz$ with $(a, b) \in I$ and $a < b$, there exists a letter c in y which is dependent from a.*

From this proposition, we deduce that the set $LexNF \subseteq \Sigma^*$ of words in lexicographic normal form is given by the star-free expression:

$$LexNF = \Sigma^* \setminus \bigcup_{(a,b)\in I, a<b} \Sigma^* b I(a)^* a \Sigma^*$$

where $I(a)$ denotes the set of letters independent from a.

Theorem 1 (Ochmanski) *Let $T \subseteq \mathbb{M}(\Sigma, I)$. The following assertions are equivalent:*

1. *T is a recognizable language.*
2. *$\varphi^{-1}(T) \cap LexNF$ is a regular language of Σ^*.*
3. *T is described by a star-connected rational expression, i.e. where the Kleen-star is used over connected traces only.*

This combinatorial proof is omitted here but can be found in [1]

2.3 Some results from logic

Recall that given the following:

1. A countable set of variables Var.
2. A countable set of constants Cst.
3. A countable family $Fct = \cup_{i \geqslant 0} Fct_i$ where Fct_i is the set of functions of arity i.
4. A countable family $Pred = \cup_{i \geqslant 0} Pred_i$ where $Pred_i$ is the set of predicates of arity i.

we define terms by induction as follows: a variable or a constant is a term, and if $f \in Fct_i$ and $t_1, \ldots, t_i$ are terms, then $f(t_1, \ldots, t_i)$ is a term. Atoms are formed in the same way using predicates.

First-order formulae are defined by induction as follows: an atom is a formula, and if A, B are formulae and $x \in Var$, the following are formulae: $\neg A$, $A \Rightarrow B$, $A \wedge B$, $A \vee B$, $\exists x A$ and $\forall x A$.

Using first-order formulae, we define a predicate calculus called first-order logic (FO) using the axiom schemata of propositional calculus together with following:

1. Quantifiers equivalence: $\exists x F \Rightarrow \neg \forall x \neg F$ and $\forall x F \Rightarrow \neg \exists x \neg F$.
2. Instantiation $\forall x F(x) \Rightarrow F[x \leftarrow r]$.
3. Limited inversion $\forall x (F \Rightarrow G) \Rightarrow (F \Rightarrow \forall x G)$ if x is not free in F.

The inference rules are modus ponens ($A, A \Rightarrow B \vdash B$) and generalization ($A \vdash \forall x A$).

Example 2 (Presburger arithmetic) *Presburger arithmetic is the first-order theory with equality (along with its axioms) having constants $\{0, 1\}$, a unary successor function, a binary function $+$ and a binary predicate $<$. Axioms are the universal closure of the following:*

1. $\neg(0 = x + 1)$
2. $x + 1 = y + 1 \Rightarrow x = y$
3. $x + 0 = x$
4. $(x + y) + 1 = x + (y + 1)$
5. *If $P(x)$ is a first-order formula in Presburger arithmetic having a free variable x, the following axiom schema holds:* $(P(0) \wedge \forall x (P(x) \Rightarrow P(x + 1))) \Rightarrow P(y)$

This theory has a natural and more importantly decidable interpretation over the integers, as we will see below.

Second-order logic is obtained by adding two countable sets of predicate variables and function variables, which can be used with quantifiers. Monadic second-order logic, denoted MSO, is the fragment of second-order logic without function variables and where predicate variables are monadic (they can only be used to quantify unary predicates).

FO and MSO have a nice interpretation in formal languages. The domain of the interpretation is Σ^*. First-order variables denote positions in a word, while predicate variables denote sets of positions. We will use a position order predicate $<$, a succesor predicate $S(x, y)$ which is true when y is the position following x and a set of predicates $(Q_a)_{a\in\Sigma}$ such that $Q_a(x)$ is true when there is an a at position x.

Using this interpretation, we denote $\mathcal{L}(\psi)$ the language defined by a closed formula ψ (i.e the set of words that satisfies it).

Example 3 *Let the FO sentence:*

$$\psi \equiv \forall x \forall y (Q_a(x) \wedge Q_a(y)) \Rightarrow \exists z (x < z \wedge z < y \wedge Q_b(z))$$

The language defined by ψ is the set of words such that any two consecutive a's have at least one b inbetween:

$$\mathcal{L}(\psi) = \Sigma^* \setminus (\Sigma^* a (\Sigma \setminus \{b\})^* a \Sigma^*)$$

Theorem 2 *A language is defineable by an MSO sentence if and only if, it is recognizable.*

Proof: ($\Leftarrow$): Assume that $\mathcal{A} = (Q, I, \delta, F)$ is a finite state automaton that recognizes L. We give a MSO formula to express that $\mathcal{A}$ has an accepting run. If $Q = \{q_1, \ldots, q_n\}$, this formula is $\exists X_1 \cdots, \exists X_n, (\psi_i \wedge \psi_t \wedge \psi_f)$, where

$$\psi_i \equiv \forall x \left(first(x) \Rightarrow \left(\bigwedge_{j \leqslant n} x \in X_j \Rightarrow \bigvee_{q_i \in I, (q_i, a, q_j) \in \delta} Q_a(x) \right) \right)$$

$$\psi_t \equiv \forall x \forall y \left(S(x, y) \Rightarrow \left(\bigwedge_{j \leqslant n} y \in X_j \Rightarrow \bigvee_{(q_i, a, q_j) \in \delta} x \in X_i \wedge Q_a(y) \right) \right)$$

$$\psi_f \equiv \forall x \left(last(x) \Rightarrow \bigvee_{q_j \in F} x \in X_j \right)$$

where $first(x) = \neg(\exists y, S(y, x))$ and $last(x) = \neg(\exists y, S(x, y))$.

($\Rightarrow$): For the converse, we only give a sketch of proof inspired from [3]. We extend our interpretation to $(\Sigma \times 2^{V_1} \times 2^{V_2})^*$ such that a letter (a, U_1, U_2) in a word w satisfied by a formula ψ (possibly with free variables) is such that all

first-order variables that refer to the position of a are in U_1 and second-order variables which contain the position of a are in U_2. We can then proceed by induction on ψ. Because of closure properties of regular languages, only the cases $\exists x\psi'$ and $\exists X\psi'$ matter. By induction hypothesis, the language defined by ψ' is recognized by a NFA $\mathcal{A} = (\Sigma \times 2^{V_1} \times 2^{V_2}, Q, q_0, \delta, F)$, from which we build a DFA $\mathcal{A}' = (\Sigma \times 2^{V_1} \times 2^{V_2}, Q \times \{0,1\}, (q_0, 0), \delta', F \times \{1\})$ where transitions are of the two following form:

$$((q,u),(a,U_1,U_2),(q',u)) \in \delta'$$

where $u \in \{0,1\}$, $x \notin U_1$ and $(q,(a,U_1,U_2),q') \in \delta$; and

$$((q,0),(a,U_1 \setminus \{x\},U_2),(q',1)) \in \delta'$$

where $x \in U_1$ and $(q,(a,U_1,U_2),q') \in \delta$. Because of our modified interpretation, it is easy to see that $\mathcal{L}(\mathcal{A}') = \mathcal{L}(\exists x\psi')$. Using a similar construction using U_2 instead of U_1, it can be shown that $\mathcal{L}(\exists X\psi')$ is also regular. $\diamondsuit$

To extend this result to traces, we introduce dependence graphs which have a natural MSO interpretation.

Definition 6 *Let (Σ, D) be a dependence alphabet. A dependence graph is (an isomorphism class of) a node-labeled acyclic graph $[V, E, \lambda]$ where*

- *V is a finite set of vertices*
- *$E \subseteq V \times V$ is an edge relation such that (V, E) is acyclic*
- *$\lambda : V \to \Sigma$ is a node labeling such that $(\lambda(x), \lambda(y)) \in D$ if and only if $(x, y) \in E \cup E^{-1} \cup id_V$.*

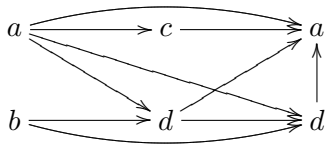
The set of dependence graphs $\mathbb{G}(\Sigma, D)$ is a monoid for the following product:

$$[V_1 \uplus V_2, E_1 \uplus E_2 \uplus \{(x,y) \in V_1 \times V_2 \mid (\lambda_1(x), \lambda_2(y)) \in D\}, \lambda_1 \uplus \lambda_2]$$

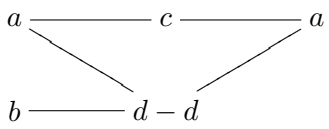
where the unit element is the empty graph.

From a trace $t = a_1 \cdots a_n \in \mathbb{M}(\Sigma, I)$, it is easy to build a dependence graph $[V = \{1, \ldots, n\}, E, \lambda] \in \mathbb{G}(\Sigma, D)$. We set $\lambda(i) = a_i$, and $(i,j) \in E \Leftrightarrow i < j \wedge (a_i, a_j) \in D$. In fact, $\mathbb{M}(\Sigma, I) \cong \mathbb{G}(\Sigma, D)$.

Example 4 *Using the same independence alphabet as in example 2.1, the dependence graph of the trace $[abcdda]_I$ is the following:*



For clarity and convenience, we often omit edges which can be obtained by transitivity. We may also omit the direction of edges. In this case, the graph reads from left to right. For instance, the reduced representation of the above dependence graph (also called Hasse diagram) is:



We now consider a MSO fragment with the following predicates:

$$(x, y) \in E, \lambda(x) = a \text{ with } a \in \Sigma$$

and their obvious interpretation in $\mathbb{G}(\Sigma, D)$. Using the correspondence between traces and dependence graphs, we can define the trace language defined by a sentence.

Theorem 3 *[4] Let $T \subseteq \mathbb{M}(\Sigma, I)$ and $K \subseteq LexNF$.*

1. *K is MSO-defineable over Σ^* if and only if, $[K]_I$ is MSO-definable over $\mathbb{M}(\Sigma, I)$.*
2. *T is MSO-defineable over $\mathbb{M}(\Sigma, I)$ if and only if, $\varphi^{-1}(T)$ is MSO-definable over Σ^*.*

Proof: 1. If $t = [V, E, \lambda]$ is a trace and $LNF(t) = a_1 \cdots a_n \in LexNF$, each node $x \in V$ corresponds to one a_i in $LNF(t)$. We will give a FO formula $\mathrm{lex}(x, y)$ that holds if and only if x is before y in lexicographic order. Once such a formula is known, if ψ defines K, we can replace all predicates $x < y$ in ψ with $lex(x, y)$. We obtain a formula $\hat{\psi}$ such that $w \models \psi \Leftrightarrow \varphi(w) \models \hat{\psi}$. The formula $\mathrm{lex}(x, y)$ can be recursively expressed:

$$\begin{aligned} \mathrm{lex}(x, y) \quad = \quad & (x, y) \in E^+ \\ & \vee(\lambda(x) < \lambda(y) \wedge \neg \mathrm{lex}(y, x)) \\ & \vee(\lambda(y) < \lambda(x) \wedge \exists z, \lambda(x) < \lambda(z) \wedge \mathrm{lex}(x, z) \wedge (z, y) \in E^*) \end{aligned}$$

Where $(x, y) \in E^+$ can be expressed as:

$$(x, y) \in E \vee \bigvee_{k \leqslant |\Sigma| - 2} \left(\exists z_1 \cdots z_k, (x, z_1) \in E \wedge \left(\bigwedge_{1 < i \leqslant k} (z_{i-1}, z_i) \in E \right) \wedge (z_k, y) \in E \right)$$

But since the alphabet Σ is finite, we can unfold the recursion to obtain the desired formula for $\mathrm{lex}(x, y)$.

2. If T is MSO-defineable, we can replace all predicates $(x, y) \in E$ by the FO formula $x < y \wedge (\lambda(x), \lambda(y)) \in D$ and all predicates $\lambda(x) = a$ by $Q_a(x)$. We obtain a sentence that defines the set of linearizations of traces in T. Conversely, because of Ochmanski's theorem, $T = \varphi(\varphi^{-1}(T) \cap LexNF)$ and (1.) applies. $\diamondsuit$

2.4 Rational trace languages

Recall that in free monoids, Kleene's theorem states that $\mathrm{Rec}(M) = \mathrm{Rat}(M)$. This result does not hold in all non-free monoids. An important counter-example of this fact is given on the monoid $a^* \times b^*$ by the language $(ab)^*$: its commutative closure is the non-regular language $\{w \in \{a,b\}^* \,|\, |w|_a = |w|_b\}$.

Because regular expressions provide the easiest description of recognizable languages, one can wonder if it can be decided whether a regular expression over a monoid represents a recognizable language or not. This question has been answered in [5] by Sakarovitch in the case of trace monoids.

Theorem 4 *Recognizability is decidable in* $\mathrm{Rat}\,\mathbb{M}(\Sigma, I)$ *if and only if,* I *is transitive.*

Although we will not give a complete proof of this theorem, some noteworthy intermediate results deserve to be stated here, as they are related to some of the problems we address in section 5.

Definition 7 *If* M *and* N *are two monoids, the* free product *of* M *and* N*, denoted* $M * N$*, is the set of finite sequences* $(u_i)_{i \leqslant n}$ *of elements alternating from* $M^\bullet$ *to* $N^\bullet$*, i.e. such that* $u_i \in M^\bullet \Leftrightarrow u_{i+1} \in N^\bullet$*, where* $M^\bullet$ *stands for* $M \setminus \{1_M\}$*.*

We define alternating concatenation by induction as follows: $(u_1, \ldots, u_p) \cdot (v_1, \ldots, v_q)$ *is equal to:*

1. $(u_1, \ldots, u_p, v_1, \ldots, v_q)$ *if* u_p *and* v_1 *belong to different monoids.*
2. *Else,* $(u_1, \ldots, u_p v_1, \ldots, v_q)$ *if* $u_p v_1 \neq 1$*.*
3. $(u_1, \ldots, u_{p-1}) \cdot (v_2, \ldots, v_q)$ *otherwise.*

$M * N$ *is a monoid for alternating concatenation where the unit element is the empty sequence.*

Lemma 1 *If* I *is transitive, then* $\mathbb{M}(\Sigma, I)$ *is isomorphic a free product of free commutative monoids.*

Proof: If I is transitive, connected components of the independence graph are complete subgraphs and the induced connected subalphabets yield free commutative monoids. Traces in $\mathbb{M}(\Sigma, I)$ can then be written as a dependent product of commutative factors. Hence, $\mathbb{M}(\Sigma, I)$ is isomorphic to the free product of free commutative monoids over those subalphabets. $\diamondsuit$

Because of this lemma, we can prove theorem 4 by showing it holds on free commutative monoids, and that it is preserved by taking free products. We will focus on the base case, which is attributed to Ginsburg and Spanier [6], although this result is not explicitely given in their paper.

Proposition 3 *Recognizability is decidable in* $\mathrm{Rat}\,\mathbb{M}$ *when* $\mathbb{M}$ *is the free commutative monoid over* Σ.

Proof: It suffices to show that the set of lexicographic forms, which are included in $a_1^* a_2^* \cdots a_n^*$ if $\Sigma = \{a_1, \ldots, a_n\}$, is regular. To a regular expression over $\mathbb{M}$, one can associate through the Parikh mapping, a semilinear set of $\mathbb{N}^n$, which are given two finite sets $C, P \subset \mathbb{N}^n$ of constants and periods, the set of elements of the form $x = x_0 + \cdots + x_m$ with $x_0 \in C$ and $(x_i)_{i \leqslant m}$ is a finite sequence of periods.

It is proved in [6] that semilinears sets are exactly the sets of $\mathbb{N}^n$ that can be defined by a Presburger formula over an adequate interpretation. But the recognizable semilinear sets are those whose periods have exactly one non-zero component, a property that is easily defined in Presburger arithmetic. $\diamondsuit$

3 Varieties of semigroups and languages

Only a limited introduction and a few results used later on are presented in this section, which owes a lot to the survey by Jean-Éric Pin in [7]. The theory of varieties originates in a renowned result by Schützenberger, who characterized star-free languages using their syntactic monoids, namely, those that are finite and aperiodic (i.e. group-free). This is in fact an instance of a general correspondence between varieties of languages and varieties of semi-groups.

3.1 Algebraic definitions

A *quasi-order* is a reflexive and transitive relation. If $\mathcal{R}$ is a quasi-order, we can define an associated equivalence relation $\mathcal{S}$ by $x \mathcal{S} y \Leftrightarrow x \mathcal{R} y \wedge y \mathcal{R} x$. We say that a relation $\mathcal{R}_1$ is coarser than $\mathcal{R}_2$ if $\mathcal{R}_2 \subseteq \mathcal{R}_1$.

A semigroup is a set with an internal associative operation. In particular, monoids are semigroups with an identity element. If S is a semigroup, we define S^1 as S if S is a monoid and $S \uplus \{1\}$ else, where 1 is a unit element.

An *ordered* semigroup is a semigroup with a stable order relation $\leqslant$ (i.e. compatible with the internal operation). A morphism of ordered semigroups $\varphi : (S_1, \leqslant_1) \to (S_2, \leqslant_2)$ is a morphism of semigroups compatible with the order relation: $\forall x, y \in (S_1, \leqslant_1), x \leqslant_1 y \Rightarrow \varphi(x) \leqslant_2 \varphi(y)$.

Recall that a congruence in a semigroup is a stable equivalence relation. If S is a semigroup and $\sim$ is a congruence, $S/\!\sim$ is also a semigroup. A congruence in an ordered semigroup $(S, \leqslant)$ is a stable quasi-order $\preceq$ that is coarser than $\leqslant$. If $\preceq$ is a congruence and $\sim$ the associated equivalence relation, then $(S/\!\sim, \leqslant)$ is also an ordered semigroup.

If φ is a morphism of (ordered) semigroups, the equivalence relation (quasi-order) defined by $x \sim_\varphi y \Leftrightarrow \varphi(x) = \varphi(y)$ (resp. $x \preceq_\varphi y \Leftrightarrow \varphi(x) \leqslant \varphi(y)$) is a congruence.

An order ideal of an ordered semigroup $(S, \leqslant)$ is an ideal $I \subseteq S$ of the semigroup S ($S^1IS^1 = I$) such that if $x \leqslant y$ and $y \in I$ then $x \in I$. The order ideal generated by an element x is the set $x\downarrow$ of all $y \leqslant x$. If S is a monoid, $K \subseteq S$ and I is an order ideal, then $I^{-1}K$ and KI^{-1} are also order ideals.

An element $e \in S$ is an idempotent if $e^2 = e$. If S is a finite semigroup and $s \in S$, since the subsemigroup generated by s is also finite, there exists a unique idempotent power of s which we denote s^ω, such that s^ω is idempotent.

Definition 8 *Let $\varphi : (S_1, \leqslant_1) \to (S_2, \leqslant_2)$ be a morphism of ordered semigroups. We say that $Q \subseteq S_1$ is recognized by φ if there exists an order ideal $P \subseteq S_2$ such that $Q = \varphi^{-1}(P)$.*

Notice for instance that the congruence $\leqslant_1$ recognizes all order ideals of S_1. If $Q \subseteq S_1$ is an order ideal, we define the syntactic congruence $\preceq_Q$ by:

$$u \preceq_Q v \iff \forall x, y \in S, (xvy \in Q \Rightarrow xuy \in Q)$$

It is the coarsest congruence of ordered semigroups that recognizes Q. The quotient $T/\preceq_Q$ is an ordered semigroup called the ordered syntactic semigroup.

All recognizable languages can be recognized by an ordered semigroup. If (Q, δ, F) is an automaton that recognizes L, the syntactic congruence $\preceq_F$ defined above is an order over the semigroup Q.

3.2 Varieties of semigroups

Definition 9 *A variety of semigroups is a class of semigroups closed under taking subsemigroups, quotients and direct products. A pseudovariety is a variety of finite semigroups.*

Varieties can be defined using identities. If Σ is an alphabet and $u, v \in \Sigma^+$, we say that a semigroup S satisfies the identity $u = v$ ($u \leqslant v$) if for every morphism of (ordered) semigroups $\varphi : \Sigma^+ \to S$ one has $\varphi(u) = \varphi(v)$ ($\varphi(u) \leqslant \varphi(v)$). It is relatively easy to see that identities define varieties, but he converse is also true:

Theorem 5 *[8] A class of (ordered) semigroups is a variety if and only if it can be defined by a set of identities.*

For instance, $xy = yx$ defines the variety of commutative semigroups. Because we are mainly interested in recognizable languages, we would like a similar theorem for pseudovarieties. Such a theorem was given by Reiterman but as we shall see, it requires some work:

Theorem 6 *A class of (ordered) semigroups is a pseudovariety if and only if it can be defined by a set of identities of $\hat{\Sigma}^+$.*

The difficulty of this theorem comes from the definition of $\hat{\Sigma}^+$. We say that a semigroup S separates two words $u, v \in \Sigma^+$ if there exists some semigroup morphism $\varphi : \Sigma^+ \to S$ such that $\varphi(u) \neq \varphi(v)$. We now define:

$$r(u, v) = \min\{|S| \mid \text{S is a finite semigroup that separates u and v}\}$$

It can be shown that $d(u, v) = 2^{-r(u,v)}$ is a metric that makes (Σ^+, d) a topological semigroup. We then define $\hat{\Sigma}^+$ as the completion of (Σ^+, d). If we consider finite semigroups as topological semigroups for the discrete metric, we say that a finite semigroup S satisfies the identity $u = v$ $(u \leqslant v)$ whith $u, v \in \hat{\Sigma}^+$, if for any continuous morphism of (ordered) semigroups $\varphi : \hat{\Sigma}^+ \to S$, $\varphi(u) = \varphi(v)$ $(\varphi(u) \leqslant \varphi(v))$.
Beside words from Σ^+, $\hat{\Sigma}^+$ contains limits of Cauchy sequences of words. An important and useful example is given below:

Example 5 *For all $x \in \hat{\Sigma}^+$, $(x^{n!})_n$ is a Cauchy sequence and its limit, denoted x^ω, is an idempotent of $\hat{\Sigma}^+$.*

In particular, if $\varphi : \hat{\Sigma}^+ \to S$ is a continuous morphism onto a finite semigroup, $\varphi(x^\omega)$ is equal to the unique idempotent power $\varphi(x)^\omega$ of $\varphi(x)$. For instance, the identity $x^\omega y^\omega = y^\omega x^\omega$ defines the pseudovariety of semigroups where idempotents commute.

3.3 Varieties of languages

A class of recognizable languages is a map $\mathcal{C}$ which associates to a finite alphabet A a set $\mathcal{C}(A^+)$ of recognizable languages over A^+. In this section, we invastigate the relationship between pseudovarieties of semigroups and classes of recognizable languages, using the corresponance $\mathbf{V} \to \mathcal{V}(A^+)$ given by the syntactic semigroup.

Definition 10 *A positive variety of languages is a class $\mathcal{V}$ of recognizable languages such that:*

1. *For every finite alphabet A, $\mathcal{V}(A^+)$ is a positive boolean algebra*
2. *For all semigroup morphism $\varphi : A^+ \to B^+$, $L \in \mathcal{V}(B^+)$ implies $\varphi^{-1}(L) \in \mathcal{V}(A^+)$.*
3. *If $L \in \mathcal{V}(A^+)$ and $a \in A$, then $a^{-1}L, La^{-1} \in \mathcal{V}(A^+)$.*

A variety of languages is a positive variety closed under complementation.

The following theorem, due respectively to Eilenberg and Pin [9] states the relationship between varieties of languages and varieties of semigroups:

Theorem 7 *$\mathbf{V} \to \mathcal{V}$ is a one to one correspondence between varieties of (ordered) finite semigroups and (positive) varieties of languages.*

Example 6 *The variety of finite aperiodic semigroups defined by the identity* $x^\omega = x^{\omega+1}$ *corresponds to the variety of star-free languages.*
The variety of finite commutative groups **Gcom** *defined by the identities* $xy = yx$ *and* $x^\omega y = yx^\omega = y$ *corresponds to the boolean algebra of languages generated by the family:*

$$F(a, k, n) = \{u \in A^* \mid |u|_a \equiv k \mod n\}$$

The variety of p-groups $\mathbf{G}_p$ *defined by the identity* $x^(p^\omega) = 1$ *corresponds [10] to the boolean algebra of languages generated by the family:*

$$S(u, k, p) = \{w \in A^* \mid \binom{w}{u} \equiv k \mod p\}$$

for $u \in A^*$, $0 \leqslant k < p$.

3.4 The positive variety $\mathcal{W}$

This positive variety of language introduced in [11] is relevant to our topic because we'll show it is closed under total commutation.

Definition 11 *The polynomial closure of a class of language* $\mathcal{L}$ *over* Σ *is the set of languages that are finite unions of languages with the form* $L_0 a_1 L_1 a_2 \cdots a_n L_n$ *whith* $a_i \in \Sigma$ *and* $L_i \in \mathcal{L}$.

We denote by $Pol(\mathcal{L})$ the polynomial closure of the variety of languages $\mathcal{L}$. It is a positive boolean algebra closed under quotients, product, shuffle and inverses of morphisms.

Definition 12 *The variety of ordered monoids* **W** *is defined as follows:* $(M, \leqslant) \in \mathbf{W}$ *if and only if, for every pair* (a, b) *of mutually inverse elements of* M *(i.e. such that* $aba = a$ *and* $bab = b$*) and any element* z *of the minimum ideal of the submonoid generated by* a *and* b*,*

$$(abzab)^\omega \leqslant ab$$

We denote $\mathcal{W}$ the positive variety of languages associated to **W**. Although there is no combinatorial description of languages in $\mathcal{W}$, it is the unique maximal positive variety that does not contain the language $(ab)^*$ with $a \neq b$. It is closed under quotiens, residuals, inverse of morphisms, length-preserving morphisms, shuffle and as we will see now, total commutation closure, and contains $\mathcal{C}om$ the variety of commutative languages and $Pol(\mathcal{G})$.

4 Commutative closures in $\mathcal{G}$ and $\mathcal{W}$

This section summarizes the latest results presented by Gómez, Guaiana and Pin at ICALP'08 [12] regarding the closure of $\mathcal{W}$ and $Pol(\mathcal{G})$ under total commutation and partial commutation closures of polynomials of group languages for some independence relations. Understanding and extending those results was the main goal of the internship.

The proofs of these results often require some variants of Ramsey's theorem, which we recall below in a very generic form:

Theorem 8 *In any finite coloring of a sufficiently large complete graph, there is a monochromatic complete induced subgraph.*

4.1 Group languages

In this section, L is a group language and $\varphi : \Sigma^* \to G$ is the projection over the syntactic group of L. We will need the following consequence of Ramsey's theorem:

Lemma 2 *For every $n > 0$, there exists some $N > 0$ such that for all $u_0, \ldots, u_N \in \Sigma^*$, there exists a sequence $0 \leqslant i_0 < i_1 < \cdots < i_n \leqslant N$ such that $\varphi(u_{i_0} u_{i_0+1} \ldots u_{i_1-1}) = \cdots = \varphi(u_{i_{n-1}} u_{i_{n-1}+1} \ldots u_{i_n-1}) = 1$.*

Theorem 9 *The commutative closure of a group language is regular.*

Proof: To show that a commutative language is regular, it suffices to prove that each letter of the alphabet is of finite index for the syntactic equivalence $\sim_{[L]}$. Let $n = |G|$, $a \in \Sigma$, $g = \varphi(a)$ and N the integer given by the previous lemma. We claim that $a^N \sim_{[L]} a^{N+n}$:

If $xa^N y \in [L]$, it is commutatively equivalent to some $w \in L$. Since $g^n = 1$, $\varphi(wa^n) = \varphi(w)$, hence $wa^n \in L$. But wa^n is commutatively equivalent to $xa^{N+n}y \in [L]$.

Conversely, if $xa^{N+n}y \in [L]$, it is commutatively equivalent to a word $w = w_0 a w_1 a \cdots w_N a w_{N+1} \in L$. We apply the previous lemma to $(w_i a)_{0 \leqslant i \leqslant N}$, yielding a sequence $0 \leqslant i_0 < i_1 < \cdots < i_N \leqslant N$ such that for all $0 \leqslant j < N$, $\varphi(f_j) = 1$ where $f_j = \prod_{i_j \leqslant k < i_{j+1}} w_k a$. Let $g_j = f_j a^{-1}$, s and t such that $w = s(\prod_{0 \leqslant j < N} f_j)t$ and $w' = s(\prod_{0 \leqslant j < N} g_j)t$. We have $\varphi(w) = \varphi(s)\varphi(t)$ and $\varphi(w') = \varphi(s)\varphi(a)^{-n}\varphi(t) = \varphi(w)$, subsequently $w' \in L$. But w' is commutatively equivalent to $xa^N y \in [L]$. ♢

A consequence of this result is that $Pol(\mathcal{G})$ is closed under total commutation, while $\mathcal{G}$ is not. Taking the polynomial closure usually increases robustness with respect to commutative closures. For instance, the varieties of piecewise testable

languages and commutative languages $\mathcal{J}$ and $\mathcal{C}om$ are closed under total commutation, while their polynomial closures are closed under partial commutation.

Given this result, one may wonder whether $Pol(\mathcal{G})$ is closed under partial commutation. A partial answer is given by Gómez, Guaiana and Pin while we provide some new evidence in section 5. First, the answer is known for a special case of dependence alphabet:

Theorem 10 *If (Σ, D) is transitive, and L is a polynomial of group languages, then $[L]_I \in Pol(\mathcal{G})$.*

In this case, $\mathbb{M}$ is a direct product of free monoids, and one can show that the closure $[L]_I$ is a finite union of direct products of polynomials of group languages over $\mathbb{M}$. $\bar{L}^I$ is then the finite union of the shuffle product of those languages. Finally, a last theorem regarding $Pol(\mathcal{G})$ is given. It is in fact a consequence of the above theorem and theorem 13 below. It applies to the case when I does not contain any of the following two induced subgraphs:

Theorem 11 *If I is (P_4, Paw)-free and $L \in Pol(\mathcal{G})$, then $[L]_I$ is recognizable.*

4.2 Languages of $\mathcal{W}$

The main result of this section is the closure of $\mathcal{W}$ under total commutation. In fact, it suffices to show that the closure of any language in $\mathcal{W}$ is regular, since $\mathcal{C}om \subset \mathcal{W}$. However, the proof of this claim is slightly subtle, since the only description of $\mathcal{W}$ is given by the pseudoidentities $(abzab)^\omega \leqslant ab$ for semigroups in $\mathbf{W}$.

In this section, we fix $L \in \mathcal{W}(\Sigma^*)$ and $\varphi : \Sigma^* \to M$ is the syntactic projection, with $M \in \mathbf{W}$.

We will need another consequence of Ramsey's theorem:

Lemma 3 *Let $a \in \Sigma$. For any $n \geqslant 0$, there exists some $N(n)$ such that for all $u \in \Sigma^*$, if $|u|_a \geqslant N(n)$, i.e. $u = u_0 u_1 a \cdots u_n a u_{N(n)+1}$, there exists an idemportent $e \in M$ such that for all $1 \leqslant i \leqslant n$, $\varphi(u_i a) = e$.*

And the following lemma for 2-letters alphabets:

Lemma 4 *Let $A = \{a, b\}$. There is a word $z \in A^*$ such that $|z|_a = |z|_b$ and for any morphism $\gamma : A^* \to M$, $\gamma(z)$ belongs to the minimal ideal of $\gamma(A^*)$.*

Proof: We define $n = |M|$ and z a word that contains every word of length $\leqslant n$ as a factor such that $|z|_a = |z|_b$. Let J_M denote the minimal ideal of $\gamma(A^*)$. If $m \in J_M$, m is the image under γ of a factor u such that $|u| \leqslant n$. Hence, u is a factor of z and $\gamma(z) \in M\gamma(u)M = J_M$. $\diamondsuit$

Theorem 12 *$[L]$ is regular.*

Proof: Let ω be the least integer such that for all $x \in M, x^\omega$ is idempotent. We will show that for all $a \in \Sigma$, there exists some $N > 0$ such that $a^{N+\omega} \leqslant_{[L]} a^N$. It is a sufficient condition for regularity in free commutative monoids (see [13]). We set the following: $n = |M|$, $z \in A^*$ is the word given by lemma 4, $r = |z|_a = |z|_b$, $n_3 = \omega(1 + r)$, $n_2 = nn_3$, $n_1 = 3n_2$ and $N = N(n_1)$ is the bound given by lemma 3.

Assume that $xa^N y \in [L]$. There exists some $u \in L$ such that $u \sim xa^N y$, so $|u|_a \geqslant N$ and there is a factorization $u = u_0(\prod_{1\leqslant i\leqslant n_1} u_i a)u_{n_1+1}$ with $\varphi(u_i a) = e$. Now, since $n_1 = 3n_2$, we set for $1 \leqslant i \leqslant n_2$ $f_i = u_{3i-2}au_{3i-1}$ and $g_i = au_{3i}a$, so that $u = u_0 \prod_{1\leqslant i\leqslant n_2}(f_i g_i)u_{n_1+1}$. Notice that $\varphi(f_i g_i) = e$, hence $\varphi(f_i g_i f_i) = e^2\varphi(u_{3_i-1}) = \varphi(f_i)$. Similarly, $\varphi(g_i f_i g_i) = \varphi(g_i)$ and $\varphi(f_i), \varphi(g_i)$ are mutually inverse.

Since $n_2 = nn_3$, by the pigeonhole principle, there is a sequence $i_1 < \cdots < i_{n_3}$ and some $s \in M$ such that for all $1 \leqslant j \leqslant n_3$, $\varphi(f_{i_j}) = s$. We set $x_j = f_{i_j}$ and $y_j = g_{i_j}$ for $1 \leqslant j \leqslant n_3$, and $\bar{s} = \varphi(a)e = \varphi(g_i)$ for all $1 \leqslant i \leqslant n_2$ (hence $s\bar{s} = e$).

By isolating the indices i_j in the factorization of u, we obtain a new factorization:

$$u = w_0 \prod_{j=1}^{n_3} x_j y_j w_j$$

with $\varphi(w_j) = e$ for $j \notin \{0, n_3\}$ and $\varphi(x_j) = s$, $\varphi(y_j) = \bar{s}$ for all j.

We now define the sequence $(z_k)_{k\leqslant\omega}$ as follows: z_k is obtained from z by replacing the ith occurence of a in z by $x_{\omega+(k-1)r+i}$ and the ith occurence of b in z by $y_{\omega+(k-1)r+i}$, for $1 \leqslant i \leqslant r$. This is legitimate since $n_1 = \omega(1 + r)$ and $|z|_a = |z|_b$. Notice that $\prod_{j=1}^{\omega} z_j \sim \prod_{j=\omega+1}^{n_3} x_j y_j$.

We now define $x'_j = u_{3i_j-2}a^2u_{3i_j-1}$ and $y'_j = az_j u_{3i_j}a$ for $1 \leqslant j \leqslant n_3$, and:

$$u' = w_0 \prod_{j=1}^{\omega} x'_j y'_j \prod_{j=1}^{n_3} w_j$$

u' is commutatively equivalent to $xa^{N+\omega}y$, since

$$\prod_{j=1}^{\omega} x'_j y'_j \sim a^\omega \prod_{j=1}^{\omega} x_j y_j \prod_{j=1}^{\omega} z_j \sim a^\omega \prod_{j=1}^{n_3} x_j y_j$$

Let T be the submonoid of M generated by s and $\bar{s}$, and $\gamma : A^* \to T$ defined by $\gamma(a) = s$ and $\gamma(b) = \bar{s}$. Lemma 4 states that $\gamma(z) \in J_T$ and by definition of $\mathbf{W}$, since $s\bar{s} = e$, $(e\gamma(z)e)^\omega \leqslant e$ (here, ω denotes the idempotent power of $e\gamma(z)e$, but it still divides the ω defined above, hence we does not dinstinguish the two).

By construction of z_j, we have for all $1 \leqslant j \leqslant \omega$ $\varphi(z_j) = \gamma(z)$. Hence, $\varphi(x'_j y'_j) = e\bar{s}\gamma(z)e$ and one has finally:

$$\varphi(u') = \varphi(w_0)(e\bar{s}\gamma(e)e)^\omega \varphi(w_{n_3})$$

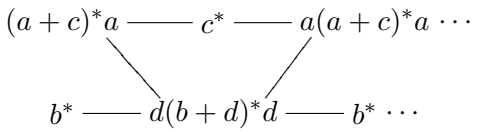
On the other hand, one can see that $\varphi(u) = \varphi(w_0)e\varphi(w_{n_3})$. Since $\bar{s} \in T$ and $\gamma(z) \in J_T$, $\bar{s}\gamma(z) \in J_T$. Using the above reduction, $\varphi(u') \leqslant \varphi(u)$ with $u \in L$, yielding $u' \in L$ and $xa^{N+\omega}y \in [L]$. $\diamondsuit$

A last theorem is given regarding the partial closure under a transitive commutation relation. Notice that unlike the total commutation case, the fact that $[L]_I$ is recognizable does not imply that it is in $\mathcal{W}$. The closure of $\mathcal{W}$ under partial commutation remains an open question.

Theorem 13 *Let $L \in \mathcal{W}(\Sigma^*)$ and I be a trasitive independence relation. Then $[L]_I$ is recognizable.*

5 Closure under a P_4 independence relation

In this section, we present the new results we obtained regarding the recognizability of I-closures of group languages when $I = P_4$. Traces in $\mathbb{M}(P_4)$ have a very specific block structure (the first two parallel blocks are optional):



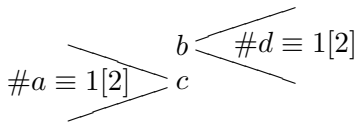
5.1 Closures of $S(u, k, n)$

Our starting point for this study was the family $S(u, k, n)$ defined above that generates the variety of p-group and nilpotent languages. We obtained two partial results for this family of languages.

Proposition 4 *$[S(abcd, 1, 2)]_I$ is regular*

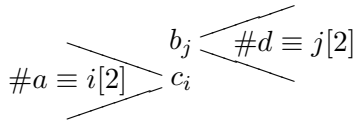
Proof: Let us start by assuming that a trace t contains the factor $\begin{matrix} b \\ c \end{matrix}$ (that is to say, a b and a c in parallel). Then, we have the following cases:

1. b is followed by an odd number of d's, and c is preceded by an odd number of a's:



   Consider the two linearizations $u_1 = xbcy$ and $u_2 = xcby$, where $|x|_a \equiv 1[2]$ and $|y|_d \equiv 1[2]$. Then, $\binom{u_1}{abcd} \equiv \binom{u_2}{abcd} + 1 \mod 2$. Regardless of how other $abcd$ subwords can be formed from the trace, $t \in [S(abcd, 1, 2)]_I$ since commuting b and c switches the parity of the number of occurences in the linearization. We call such situations *flipping cases*.

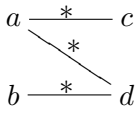
2. Otherwise, b and c are colored with the parity of following d's (respectively, preceding a's):



   Commuting b_0 and c_1, b_1 and c_0 or b_0 and c_0 doesn't affect the parity of occurences of $abcd$.

Assume now that t contains the factor $\begin{matrix} a \\ b \end{matrix}$. The following cases may occur:

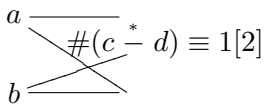
3. If t contains the subtrace:



   in other words $t = t_1\binom{a}{b}t_2\binom{c}{d}t_3$. Let x_1, x_2, x_3 be any linearizations of t_1, t_2, t_3 and $u = x_1abx_2$, $v = x_2cdx_3$.

   Notice that $\binom{ucdx_3}{abcd} \equiv \binom{udcx_3}{abcd} + 1 \mod 2$ and $\binom{x_1abv}{abcd} \equiv \binom{x_1bav}{abcd} + 1 \mod 2$. Hence, this is a flipping case for any $t_1, t_2, t_3 \in \mathbb{M}$.

4. If a is followed by an odd number of occurrences of the subtrace $c \overset{*}{—} d$ (that is, a has no c and d in parallel in its future, thus we can count only subtraces cd after a such that c and d are ordered):



   we clearly have a flipping case.

5. If a is followed by an even number of occurences of the subtrace $c \overset{*}{—} d$, a's are colored using this parity and commuting a_0 and b won't change the membership to L of any linearization of t.

Finally, when t contains the factor $\begin{matrix} c \\ d \end{matrix}$, we obtain symmetric cases 6, 7 and 8 that are the counterparts of cases 3, 4 and 5.

Each of the flipping cases 1,3,4,6,7 described above is recognizable. For instance cases 1 and 3 yield the following recognizable trace languages:

$$\begin{aligned} L_1 &= \mathbb{M}_a^1\, bc\, \mathbb{M}_d^1 \\ L_3 &= \mathbb{M}\, ab\, \mathbb{M}\, cd\, \mathbb{M} \end{aligned}$$

with $\mathbb{M}_a^1$ and $\mathbb{M}_d^1$ resp., denoting the set of traces containing an odd number of a's and d's, resp. (w.l.o.g. we write recognizable trace languages instead of I-closed regular languages, which are equivalent notions).

Once the flipping cases are dealt with, we can assume we are under the conditions of cases 2, 5 and 8. Our alphabet is now $\Sigma' = \{a_0, a_1, b_0, b_1, c_0, c_1, d_0, d_1\}$, and the independence relation is (the symmetric closure of) $I' = \{(a_0, b_0), (a_0, b_1), (b_0, c_0), (b_0, c_1), (b_1, c_0), (d_0, c_0), (d_0, c_1)\}$. We have shown that $\chi(S(abcd, 1, 2) \setminus \text{Flip})$ with $\text{Flip} = L_1 \cup L_3 \cup L_4 \cup L_6 \cup L_7$, is I'-closed – where χ is the coloring defined in 2, 5, 8. Let $w \in [S(abcd, 1, 2)]_I \setminus \text{Flip}$. Thus, $w \equiv_I v$ for some $v \in S(abcd, 1, 2) \setminus \text{Flip}$. We have also $\chi(w) \equiv_{I'} \chi(v)$, thus $w \in S(abcd, 1, 2)$ too.

We obtain the claimed result:

$$[S(abcd, 1, 2)]_I = \text{Flip} \cup S(abcd, 1, 2)$$

$\diamondsuit$

Using this idea, we proved a slightly more general result: for any independence relation I and for any $u \in \Sigma^*$ such that $|u|_{ab} + |u|_{ba} \leqslant 1$ for each $(a, b) \in I$, a property we'll abbreviate to $\wp$, the language $[S(u, 1, 2)]_I$ is regular.

Proposition 5 *Let $u \in A^*$ be such that $\wp$ is satisfied. A trace t is in the I-closure of $S(u, r, 2)$ if and only if either one of the following conditions is satisfied:*

1. *There exists a factorization $u = xaby$ such that $(a, b) \in I$, and a factorization $t = t_1 abt_2$ such that $t_1 \in [S(x, 1, 2)]_I$ and $t_2 \in [S(y, 1, 2)]_I$.*
2. *$\varphi^{-1}(t) \subseteq S(u, 1, 2)$*

Proof: Clearly, the second condition is sufficient. Assuming 1, there exist some $v \in S(x, 1, 2), z \in S(y, 1, 2)$ such that $t_1 = [v]_I$ and $t_2 = [z]_I$. Let $w = vabz$ and $w' = vbaz$. Assume there is a subword $xaby$ of w that contains our highlighted (a, b) pair. Since $xaby$ is the only factorization of u containing $(a, b) \in I$, this subword can be split into a subword x of v and a subword y of z, thus

$$\binom{w}{u} \equiv \binom{w'}{u} + \binom{v}{x}\binom{z}{y} \equiv \binom{w'}{u} + 1 \mod 2$$

Notice that this sentence is false when there are multiple occurences of the pair (a, b) in u. For instance, $\binom{aaba}{aba} = \binom{abaa}{aba}$.
Conversely, assume that $t \in [S(u, 1, 2)]_I$ and suppose we exclude the first case. There exists some $v \in S(u, 1, 2)$ such that $t = [v]_I$. If $v = wabz$ is a factorization of v with $(a, b) \in I$, we claim that $wbaz \in S(u, 1, 2)$.

- If ab does not appear in u, commuting a and b in v will have no effect.
- Otherwise, let $u = xaby$ be any factorization containing ab. Since we excluded the first case, either $w \in S(x, 0, 2)$ or $z \in S(y, 0, 2)$. Either way,

$$\binom{wabz}{u} \equiv \binom{wbaz}{u} \mod 2$$

Hence, since the claim holds for any factorization containing a pair of commuting letters, $\varphi^{-1}(t) \subseteq S(u, 1, 2)$. ♢

Proposition 6 *For each $u \in \Sigma^*$ that satisfies $\wp$, $[S(u, 1, 2)]_I$ is regular.*

Proof: We proceed by induction on $|u|$. If u is a single letter, the proposition is trivial. Otherwise, proposition 2 yields

$$[S(u, 1, 2)]_I = S(u, 1, 2) \cup \bigcup_{\substack{u=xaby \\ (a,b)\in I}} [S(x, 1, 2)]_I ab [S(y, 1, 2)]_I$$

♢

Unfortunately, there doesn't seem to be a simple generalization to any $u \in A^*$ using this approach. Our attempt to index the occurences of each letters in u in decreasing order, yielding a word $\bar{u} \in B^*$ having all its letters distinct has proved unsuccessful. Using the homomorphism

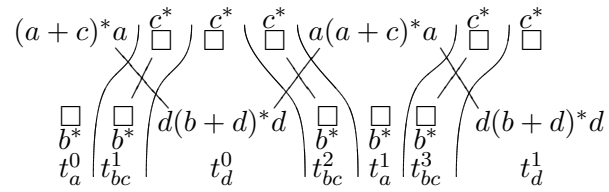
$$h : \begin{array}{ccc} A^* & \longrightarrow & B^* \\ a & \longmapsto & a_1 \cdots a_{|u|_a} \end{array}$$

it can be shown that $t \in [S(u, r, 2)]_I \Rightarrow h(t) \in [S(\bar{u}, r, 2)]_I$, but the converse does not hold: if $I = a—b$, $u = aba$ and $t = a—c—\binom{a}{b}—c—a$, $a_1a_2ca_1b_1a_2ca_1a_2$ is an odd linearization of $h(t)$ (i.e it has an odd number of occurences of the subword $\bar{u} = a_2b_1a_1$) whereas all linearizations of t are even.

5.2 P_4-closures of group languages

Using a more algebraic approach and relying on theorem 11, we proved a much more general result. If $A \subseteq \Sigma$, we will denote $\sim_A$ the commutation equivalence over the subgraph of (Σ, I) defined by A and $\mathbb{M}_A = \Sigma^*/\sim_A$. In this section, L is a group language and its syntactic group is denoted G.

Definition 13 *We call a* distribution *of t a trace factorization $t = t_a^0 t_{bc}^1 t_d^0 t_{bc}^2 t_a^1 t_{bc}^3 \dots$ in accordance with the following scheme:*



Each group of b's in parallel with an a-block and each group of c's in parallel with a d-block is distributed between the a or c block and two surrounding bc-blocks.

Proposition 7 *A trace t is in $[L]_I$ if and only if, there is a distribution of t such that:*

1. *For all a-block t_a^i, there exists some $g_i \in G$ such that $t_a^i \in [\varphi^{-1}(g_i)]_{I_d} \cap \mathbb{M}_{a,b,c}$ where $I_d = \{(a,b),(a,c)\} \cup \{(x,d), x \in \{a,b,c\}\}$.*
2. *For all d-block t_d^i, there exists some $h_i \in G$ such that $t_d^i \in [\varphi^{-1}(h_i)]_{I_a} \cap \mathbb{M}_{b,c,d}$ where $I_a = \{(b,c),(c,d)\} \cup \{(a,x), x \in \{b,c,d\}\}$.*
3. *For all bc-block t_{bc}^i, there exists some $f_i \in G$ such that $t_{bc}^i \in [\varphi^{-1}(f_i)] \cap \mathbb{M}_{b,c}$ where $[\,\cdot\,]$ denotes the closure under total commutation.*
4. *$\prod g_i f_{2i+1} h_i f_{2i+2} \in \varphi(L)$ (where missing blocks have their corresponding element set to 1_G).*

Proof: ($\Rightarrow$): Because of the structure of traces in $\mathbb{M}$, a linearization of t corresponds to a unique distribution. If $t \in [L]_I$, there exists a linearization $w \in L$, and the associated distribution satisfies condition 4. Furthermore, if s is a block of the distribution, the corresponding factor v of w is such that $\varphi(v) = s$, hence conditions 1—3 are satisfied.

($\Leftarrow$): the actual sufficient condition for a block s on a subalphabet P associated to an element $g \in G$ is whether $s \in [\varphi^{-1}(g) \cap P^*]_{(P,I)}$. If this is the case for all blocks, condition 4 implies that t can be globally linearized into a word of L. However, since we want to use the theorem for polynomials of group languages, notice that this is equivalent to $s \in [\varphi^{-1}(g)]_{\mathcal{C}(P)} \cap \mathbb{M}_P$ where $\mathcal{C}(P) = (I \cap P \times P) \cup \{(u,v) \mid u \in \Sigma \setminus P, v \in \Sigma \setminus \{u\}\}$. $\diamondsuit$

Theorem 14 *$[L]_I$ is recognizable.*

Proof: The existence of a distribution is MSO-defineable. Let $n = |G|$. We will only give some fragments instead of a full sentence. For instance, the formulae

that define an a-block and the existence of a monochromatic b-distribution in an a-block are given bellow:

$$\begin{aligned} A(x,y) &\equiv \lambda(x)=a \wedge \lambda(y)=a \wedge \\ &\quad \forall z((x,z)\in E \wedge (z,y)\in E) \Rightarrow (\lambda(z)=a \vee \lambda(z)=c) \\ A_{\text{lin}}(x_1,y_1,y_2,x_2) &\equiv \lambda(x_1)=d \wedge \lambda(x_2)=d \wedge A(y_1,y_2) \wedge \\ &\quad \text{Mono}(x_1,y_1) \wedge \text{Mono}(y_1,y_2) \wedge \text{Mono}(y_2,x_2) \wedge \\ &\quad \exists z_1,z_2 (x_1,z_1)\in E \wedge (z_2,x_2)\in E \wedge \text{Mono}(z_1,z_2) \wedge \\ &\quad \lambda(z_1)=b \wedge \lambda(z_2)=b \end{aligned}$$

Similar formulas are easy to obtain for other types of blocks. The guessing of the group element associated with a block is done with a coloring (disjoint position sets $X_1,\dots X_n$), in which it it possible to include whether $t_a^i \in [\varphi^{-1}(g_i)]_{I_d} \cap \mathbb{M}_{a,b,c}$ (because $\varphi^{-1}(g_i)$ is a group language and I_d is (P_4, Paw)-free, hence it is defineable by an MSO sentence). The condition on $\prod f_{2i}g_i f_{2i+1}h_i$ can be tested by "chaining" our coloring, i.e. by storing the color of the product of all the group elements corresponding to all the following blocks in each letter of the current block. $\diamondsuit$

6 Final words

The soundness of the algebraic framework of the theory of recognizable languages allows some impressive results like theorem 12. Theorem 14 is, compared to proposition 6, another example of how the combinatorial difficuly can be completly eliminated by focusing on the structure of the trace monoid.

A natural continuation of theorem 14 would be to consider closured under cographs (which are exactly the P_4-free graphs). Unfortunately, we lacked time to properly address this question, and it remains unanswered for now.

References

[1] Volker Diekert and Yves Métivier. *Handbook of formal languages, vol. 3: beyond words*, chapter Partial commutation and traces, pages 457–533. Springer-Verlag New York, Inc., New York, NY, USA, 1997.

[2] Jean Berstel. *Transductions and Context-Free Languages.* Teubner Verlag, 1979.

[3] Howard Straubing. *Finite automata, formal logic, and circuit complexity.* Birkhauser Verlag, Basel, Switzerland, Switzerland, 1994.

[4] V. Diekert and G. Rozenberg. *The Book of Traces.* World Scientific, Singapore, 1995.

[5] Jacques Sakarovitch. The "last" decision problem for rational trace languages. In *LATIN '92: Proceedings of the 1st Latin American Symposium on Theoretical Informatics*, pages 460–473, London, UK, 1992. Springer-Verlag.

[6] S. Ginsburg and E. H. Spanier. Semigroups, presburger formulas, and languages. *Pacific Journal of Mathematics*, 16(2):285–296, 1966.

[7] Jean-Éric Pin. *Syntactic semigroups*, pages 679–746. Springer-Verlag New York, Inc., New York, NY, USA, 1997.

[8] J. Almeida. *Finite Semigroups and Universal Algebra*. World Scientific, Singapore, 1995.

[9] Jean-Éric Pin. A variety theorem without complementation. *Russian Mathematics (Iz. VUZ)*, 39:80–90, 1995.

[10] Denis Thérien. Languages of nilpotent and solvable groups (extended abstract). In *Proceedings of the 6th Colloquium, on Automata, Languages and Programming*, pages 616–632, London, UK, 1979. Springer-Verlag.

[11] Antonio Cano Gómez and Jean-Éric Pin. Shuffle on positive varieties of languages. *Theor. Comput. Sci.*, 312(2-3):433–461, 2004.

[12] Antonio Cano Gómez, Giovanna Guaiana, and Jean-Éric Pin. When does partial commutative closure preserve regularity? In *ICALP '08: Proceedings of the 35th international colloquium on Automata, Languages and Programming, Part II*, pages 209–220, Berlin, Heidelberg, 2008. Springer-Verlag.

[13] Aldo De Luca and Stefano Varricchio. *Finiteness and Regularity in Semigroups and Formal Languages*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1999.