Uniformisation Gives the Full Strength of Regular Languages

Nathan Lhote
University of Warsaw, Poland
http://di.ulb.ac.be/verif/lhote/
nlhote@mimuw.edu.pl

Vincent Michielini
University of Warsaw, Poland
michielini@mimuw.edu.pl

Michał Skrzypczak
University of Warsaw, Poland
https://www.mimuw.edu.pl/~mskrzypczak/
mskrzypczak@mimuw.edu.pl

Abstract

Given R a binary relation between words (which we treat as a language over a product alphabet $\mathbb{A} \times \mathbb{B}$), a uniformisation of it is another relation L included in R which chooses a single word over $\mathbb{B}$, for each word over $\mathbb{A}$ whenever there exists one. It is known that **MSO**, the full class of regular languages, is strong enough to define a uniformisation for each of its relations. The quest of this work is to see which other formalisms, weaker than **MSO**, also have this property. In this paper, we solve this problem for pseudo-varieties of semigroups: we show that no nonempty pseudo-variety weaker than **MSO** can provide uniformisations for its relations.

2012 ACM Subject Classification Theory of computation → Regular languages

Keywords and phrases pseudo-variety, finite word, semigroup, uniformisation, regular language

Digital Object Identifier 10.4230/LIPIcs.MFCS.2019.61

Funding The first two authors have been supported by the European Research Council (ERC) grant under the European Union's Horizon 2020 research and innovation programme (ERC Consolidator Grant LIPA, grant agreement No. 683080). The first author has also been supported by the DeLTA project (ANR-16-CE40-0007). The last author has been supported by Poland's National Science Centre (NCN) (grant No. 2016/21/D/ST6/00491).

1 Introduction

Regular languages of finite words lie at the core of modern automata theory. The study of their properties has led to multiple fundamental discoveries. Among these discoveries was a formal introduction of the model of non-deterministic automata by Rabin and Scott [16]. Later, the results of Büchi, Elgot, and Trakhtenbrot [2, 7, 25] laid the foundations of the correspondence between automata and Monadic Second-Order (**MSO**) logic. That correspondence is now considered a golden standard, with notable extensions to other structures, like finite and infinite trees. Another breakthrough obtained over finite words was the effective characterisation of the class of star-free languages by McNaughton, Papert and Schützenberger [19, 11]. Again, this result has opened a rich area of extensions, first to infinite words [24], and later to other structures and classes of languages [23, 21, 1]. From the perspective of these results, the theory of regular languages of finite words can be seen as a test ground for novel problems and methods.

The situation is a bit different with the problem of uniformisation. This problem asks, to find an effectively definable graph of a function that is contained in a given relation R.

Thus, it can be seen as an instance of the choice axiom: the relation $R(x, y)$ admits multiple witnesses y for each argument x and our task is to choose one of them. The origins of that problem come from descriptive set theory, with the famous theorems like that of Novikov and Kondô [9, Theorem 36.12].

The problem of uniformisation was translated to the context of automata theory by Rabin [15], directly for the most complex structures – infinite trees. In that context, the relation R is given by an **MSO**-definable language over a product alphabet $\mathbb{A} \times \mathbb{B}$, and the question of uniformisation asks to find an **MSO**-definable language that realises a function from structures over $\mathbb{A}$ to structures over $\mathbb{B}$. Therefore, the question of uniformisability can be read as the problem of effective selection of witnesses – a way of making non-determinism somehow controlled. In the course of research over that problem, it was shown that **MSO**-definable uniformisation is always possible over infinite words [20, 10, 17]; but not over infinite trees [8, 4]. In parallel, a study of *sequential uniformisation* was performed in the context of games and problems of synthesis [3].

The fact that **MSO** has[1] the uniformisation property over finite words is easy to prove and is considered folklore: it is enough to choose the lexicographically minimal witness. That is probably the reason why the case of finite words was somehow ignored in the study of uniformisation problems. It was opened by a recent paper [12], where the author asks about the possibility to uniformise for certain logics weaker than **MSO**. The results of that work are not unequivocal. On the one hand, it is shown that for multiple pairs of logics $\mathbf{L}_1 \subseteq \mathbf{L}_2 \subseteq \mathbf{MSO}$, there exists a relation R definable in $\mathbf{L}_1$ with no $\mathbf{L}_2$-definable uniformisation. On the other hand, it is shown that each relation definable in $\mathbf{FO}[\,]$ (First-Order logic with only equality and letter tests) can be effectively uniformised within First-Order logic with the order predicate. This leads to an intriguing graph of logics, one (not) uniformising another. To simplify the situation, the author formulated the above question for $\mathbf{L}_1 = \mathbf{L}_2$, *i.e.* the problem whether a given logic uniformises itself. Based on the provided results, the following conjectured:

▶ **Conjecture 1** (Conjecture 1 in [12]). *Let* $\mathbf{L}$ *be a fragment of* **MSO** *such that* $\mathbf{FO}^2[\,] \subseteq \mathbf{L}$ *and* $\mathbf{L}$ *satisfies some closure properties (to be specified). If* $\mathbf{L}$ *has the uniformisation property, then* $\mathbf{L}$ *is* **MSO**.

The main result of the present paper is a positive answer to the above conjecture. The assumption that the logic $\mathbf{FO}^2[\,]$ (the two-variable fragment of First-Order logic) is contained in a given logic turned out to be unnecessary, and the relevant closure properties boil down to the standard notion of a pseudo-variety.

▶ **Theorem 2.** **MSO** *is the unique nonempty self-uniformisable pseudo-variety of semigroups.*

A class of languages corresponds to a pseudo-variety of semigroups if it is closed under Boolean combinations, left and right quotients, and pre-images under non-erasing homomorphisms. It is known that most of the classically considered logics correspond to pseudo-varieties of semigroups [18]. Since we restrict to semigroups instead of monoids (*i.e.* we require the considered homomorphisms of words to be non-erasing), this definition also captures logics with successor instead of the order. Among the notable examples of logics that do not correspond to pseudo-varieties of semigroups are the logics with modulo predicates, like $\mathbf{FO}[\leq, \mathbf{MOD}_2]$. These logics are not covered by the presented arguments. To deal with them, one would need to consider homomorphisms of words that are *length-preserving* or *length-multiplying*, see *e.g.* [14].

[1] We identify a logic with the class of languages it defines. Therefore $\mathbf{MSO} = \mathbf{REG}$, *i.e.* the class of all regular languages.

When read in terms of logics, the above result says that most of the widely considered formalisms over finite words weaker than **MSO** do not have the uniformisation property. As generally in the case of negative results, the consequences of that are more theoretical than practical: there is no hope in finding a robust formalism, that would be easier to handle than **MSO** and still retain the ability to choose unique witnesses.

The provided proof is independent from the results of [12] – instead of comparing two logics $\mathbf{L}_1 \subseteq \mathbf{L}_2$, we focus on one formalism $\mathbf{L}$ with the assumption that $\mathbf{L}$ can uniformise itself. Based on that assumption, we gradually bootstrap the expressive power of the considered formalism. It is done in a sequence of steps, showing that $\mathbf{L}$ must be able to express more and more complex properties: test the letters that appear in a given word; recognise the order in which the letters appear; *etc.* Each of these steps is based on the assumption that $\mathbf{L}$ has the uniformisation property and therefore must be sensitive to certain modifications of the input words. Thus, the proofs of the lemmas are of a similar structure.

Nevertheless, we believe that it is non-trivial and instructive to see the ways in which uniformisability guarantees the considered expressive abilities. The difficulty of these arguments lies in the fact that the assumptions about $\mathbf{L}$ speak about the algebraic properties of the semigroups recognising languages in $\mathbf{L}$, while the notion of uniformisability is a set-theoretical property of the actual languages in $\mathbf{L}$.

While the proof goes on, we have more and more tools at hand, but at the same time we need to prove stronger and stronger expressibility properties about $\mathbf{L}$. Ultimately, it turns out that $\mathbf{L}$ is able to guess evaluations with respect to arbitrary finite semigroups (*i.e.* in a sense is closed under projection) and therefore must contain all regular languages of finite words. From this perspective, the proof can be seen as a variety of concrete recipes (ranging from the least complex properties to the most complex) explaining what is the interplay between the considered expressibility property and uniformisability.

Although the results are expressed over finite words, we believe that similar arguments can be adapted to the more complex structures, like infinite words or finite trees. Therefore, finite words are used here again as a testing ground, providing new understanding that can later be transferred to richer structures.

The paper is organised as follows. Section 2 is devoted to an introduction of all relevant technical notions. Then, Sections 3.1 up to 3.7 gradually increase the expressive power of the considered class of languages $\mathbf{L}$. Finally, in Section 4 we conclude.

2 Technical background

Words and languages

We identify each natural number n with the set $\{0, \dots, n-1\}$, and we denote the set of all natural numbers by ω. An *alphabet* $\mathbb{A}$ is any finite set. A function from a natural number n to an alphabet $\mathbb{A}$ is called a *(finite) word* over $\mathbb{A}$. The natural number n is the *length* of w, and we denote it by $|w|$. For each $i \in n$ the element $w(i) \in \mathbb{A}$ is called the ith *letter* of w. We write $w = w(0) \cdot w(1) \cdots w(n-1)$, but notice that this notation is ambiguous when the alphabet is not clear in the context: if we write $w = a \cdot b$, we do not know *a priori* if we see w as a word over $\{a, b\}$, or over any other alphabet containing $\{a, b\}$. We extend this notation: if w_1 and w_2 are two words over $\mathbb{A}$, then $w_1 \cdot w_2$, the *concatenation* of w_1 and w_2, is the word w over $\mathbb{A}$ of length $|w_1| + |w_2|$ defined by $w(i) = w_1(i)$ for $i \in |w_1|$ and $w(i) = w_2(i - |w_1|)$ for $i \geq |w_1|$.

A word of length 0 is denoted by ϵ and called the *empty word*. The set of all words over $\mathbb{A}$ is denoted by $\mathbb{A}^*$, $\mathbb{A}^n$ is the set of all words of length $n \in \omega$, and $\mathbb{A}^+$ is $\mathbb{A}^* \setminus \{\epsilon\}$, *i.e.* the set of nonempty words over $\mathbb{A}$. Note that $\emptyset^* = \{\epsilon\}$ and $\emptyset^+ = \emptyset$.

In this paper, a *language* of words over an alphabet $\mathbb{A}$ is any subset[2] of $\mathbb{A}^+$. Once again, notice that a language has to be given with its alphabet. To avoid any ambiguity like the one above, we sometimes write $\langle L, \mathbb{A}\rangle$ to emphasise the choice of the alphabet.

In order to ease the reading of the paper, we will denote by $\mathbb{A}, \mathbb{A}_0, \mathbb{A}_1 \dots$ alphabets of letters $a, b, \dots, x, y \dots$ and by $\mathbb{B}, \mathbb{B}_0, \mathbb{B}_1 \dots$ alphabets of symbols $\Box, \bullet, \triangle, \dots$ Words over the alphabets of letters will be denoted by $u, v, w \dots$ while words over the alphabets of symbols will be denoted by $\pi, \sigma, \tau, \dots$

Semigroups and pseudo-varieties

In this paper, the classes of languages which we focus on correspond to *pseudo-varieties of semigroups*.

A *semigroup* is a set S provided with an associative binary operation, which we will denote by $\cdot$. We identify the semigroup $\langle S, \cdot\rangle$ with its set S. For $s \in S$ and for a natural number $n \geq 1$, s^n denotes the product $s \cdots s$, where s appears n times. An element $e \in S$ is said to be *idempotent* if $e^2 = e$.

It is considered folklore to prove that for every finite semigroup S, there exists a natural number $n \geq 1$ such that s^n is idempotent for each $s \in S$. We denote this natural number by $\sharp(S)$. When the semigroup is known from the context, we just write $\sharp$ instead of $\sharp(S)$. Notice that in the literature the symbol ω is often used instead of $\sharp$.

If S is a semigroup, then $S' \subseteq S$ is said to be a *sub-semigroup* of S if it is stable by the operation of S, that is, if for all s, t in S', $s \cdot t \in S'$. Finally, if S_1 and S_2 are two semigroups, then $S_1 \times S_2$ provided with the operation defined by $\langle s_1, s_2\rangle \cdot \langle t_1, t_2\rangle = \langle s_1 \cdot s_2, t_1 \cdot t_2\rangle$ is also a semigroup. It is called the *product* of S_1 and S_2.

Let S_1 and S_2 be two semigroups. A *homomorphism* from S_1 to S_2 is a function α from S_1 to S_2 such that for all x, y in S_1, we have $\alpha(x \cdot y) = \alpha(x) \cdot \alpha(y)$. Such a homomorphism is *surjective* (resp. *injective*, *bijective*) if so is the respective function α.

For each alphabet $\mathbb{A}$, the set $\mathbb{A}^+$ provided with the concatenation operation, is a semigroup that is known as the *free semigroup on* $\mathbb{A}$. The fact that a homomorphism $\alpha\colon \mathbb{A}^+ \to S$ must preserve the operations of the semigroups, implies that α is uniquely determined by its action on the single letters in $\mathbb{A}$.

The following variant of Ramsey's theorem is often used when working with finite semigroups.

▶ **Theorem 3** (Simon [22], see also Section II.11.1 in [13]). *Let $\mathbb{A}$ be an alphabet and α a homomorphism from $\mathbb{A}^+$ to some finite semigroup S. For each natural number $n \geq 2$, there exists a natural number $N(n)$ such that for each word w over $\mathbb{A}$ of length at least $N(n)$, there exists an idempotent e in S and a decomposition $w = u \cdot w_0 \cdots w_{n-1} \cdot v$, where for all $i \in n$, w_i is nonempty and $\alpha(w_i) = e$.*

Let L be a language of words over some alphabet $\mathbb{A}$ and let S be a finite semigroup. We say that S *recognises* L if there exists a homomorphism α from $\mathbb{A}^+$ to S and $T \subseteq S$ such that $L = \alpha^{-1}(T)$. In such a case we also say that the tuple $\langle S, \alpha, T\rangle$ *recognises* L.

A language of words over $\mathbb{A}$ is *regular* if it is recognised by some finite semigroup S. We denote the class of all regular languages by **REG**. As mentioned in the introduction, the class **REG** coincides with the class of languages definable in Monadic Second-Order logic (denoted **MSO**); however as logic is not directly involved in the presentation, we will rather use **REG** to emphasise the automata- and semigroup-based approach.

[2] It is more standard to define languages as subsets of $\mathbb{A}^*$ but then the natural algebraic structures are monoids, and not semigroups.

Let $\mathbf{L}$ be a class of languages and $\mathbf{S}$ a class of finite semigroups, we say that $\mathbf{L}$ *corresponds* to $\mathbf{S}$ if it is exactly the class of languages recognised by the semigroups S of $\mathbf{S}$. Notice that, under this assumption, $\mathbf{L}$ only contains regular languages.

We define now an important notion: the notion of pseudo-varieties.

▶ **Definition 4.** *Let* $\mathbf{S}$ *be a class of finite semigroups. We say that* $\mathbf{S}$ *is a* pseudo-variety *of (finite) semigroups if it has the following properties:*

- *if* $S_1 \in \mathbf{S}$ *and if* S_2 *is a sub-semigroup of* S_1*, then* $S_2 \in \mathbf{S}$*,*
- *if* S_1 *and* S_2 *are in* $\mathbf{S}$*, then the product semigroup* $S_1 \times S_2$ *is in* $\mathbf{S}$*,*
- *if* $S_1 \in \mathbf{S}$ *and if there exists a surjective homomorphism from* S_1 *to* S_2*, then* $S_2 \in \mathbf{S}$*.*

The following theorem is a part of the so-called Eilenberg's variety theory.

▶ **Theorem 5** (Eilenberg [6])**.** *Let* $\mathbf{L}$ *be a class of regular languages. Then the following propositions are equivalent:*

- $\mathbf{L}$ *corresponds to a pseudo-variety of semigroups;*
- $\mathbf{L}$ *has the following closure properties:*
  - $\mathbf{L}$ *is closed under Boolean operations: for each* $\langle L_1, \mathbb{A}\rangle$ *and* $\langle L_2, \mathbb{A}\rangle$ *in* $\mathbf{L}$*,* $\langle L_1 \cup L_2, \mathbb{A}\rangle$ *and* $\langle L_1^c := \mathbb{A}^+ \setminus L_1, \mathbb{A}\rangle$ *are in* $\mathbf{L}$ *(and therefore also* $\langle L_1 \cap L_2, \mathbb{A}\rangle$*),*
  - $\mathbf{L}$ *is closed under quotients: for each* $\langle L, \mathbb{A}\rangle \in \mathbf{L}$ *and each word* $w \in \mathbb{A}^+$*,* $\langle w^{-1} \cdot L, \mathbb{A}\rangle$ *and* $\langle L \cdot w^{-1}, \mathbb{A}\rangle$ *are in* $\mathbf{L}$*, where* $w^{-1} \cdot L = \{u \in \mathbb{A}^+ \mid w \cdot u \in L\}$*, and* $L \cdot w^{-1}$ *is defined symmetrically,*
  - $\mathbf{L}$ *is closed under pre-images of semigroup homomorphisms: for each alphabet* $\mathbb{A}$*, language* $\langle L, \mathbb{B}\rangle \in \mathbf{L}$*, and homomorphism* φ *from* $\mathbb{A}^+$ *to* $\mathbb{B}^+$*, the language* $\langle \varphi^{-1}(L), \mathbb{A}\rangle$ *is in* $\mathbf{L}$*.*

All the classes of languages discussed in [12] are pseudo-varieties of semigroups. The most common are **MSO** and $\mathbf{FO}[<]$ (First-Order logic with the order). Other examples include $\mathbf{FO}^2[<]$, the fragment of $\mathbf{FO}[<]$ where only two distinct variables are allowed; and $\mathbf{FO}[\mathbf{s}]$, where instead of the order one allows the successor function $\mathbf{s}$.

Uniformisation

Let $\mathbb{A}$ and $\mathbb{B}$ be two alphabets. If $a \in \mathbb{A}$ and $\square \in \mathbb{B}$ are two letters then their pair is denoted $\left(\begin{smallmatrix} a \\ \square \end{smallmatrix}\right) \in \mathbb{A} \times \mathbb{B}$. Let w, π be two words over $\mathbb{A}$ and $\mathbb{B}$ respectively, such that $|w| = |\pi|$. Then the pair $\langle w, \pi\rangle \in \mathbb{A}^* \times \mathbb{B}^*$ can be identified with $\left(\begin{smallmatrix} w \\ \pi \end{smallmatrix}\right)$, the word over the product alphabet $\mathbb{A} \times \mathbb{B}$ satisfying $\left(\begin{smallmatrix} w \\ \pi \end{smallmatrix}\right)(i) = \left(\begin{smallmatrix} w(i) \\ \pi(i) \end{smallmatrix}\right)$ for all $i \in |w|$. This vertical notation is also extended to sets of words of fixed length, for instance, $\left(\begin{smallmatrix} \{a,b\} \\ \square \end{smallmatrix}\right) = \left\{\left(\begin{smallmatrix} a \\ \square \end{smallmatrix}\right), \left(\begin{smallmatrix} b \\ \square \end{smallmatrix}\right)\right\}$ is a set of two letters in $\mathbb{A} \times \mathbb{B}$.

Let $R \subseteq (\mathbb{A} \times \mathbb{B})^+$. Based on the previous identification, R can be seen as a binary relation between words over $\mathbb{A}$ and words over $\mathbb{B}$. The *projection* of R is the set of words $w \in \mathbb{A}^+$ such that there exists a word $\pi \in \mathbb{B}^{|w|}$ with $\left(\begin{smallmatrix} w \\ \pi \end{smallmatrix}\right) \in R$. We denote this set by $\Pi(R)$.

A *uniformisation* of R is a relation $F \subseteq R$ such that $\Pi(F) = \Pi(R)$, and being functional, *i.e.* if $\left(\begin{smallmatrix} w \\ \pi_1 \end{smallmatrix}\right) \in F$ and $\left(\begin{smallmatrix} w \\ \pi_2 \end{smallmatrix}\right) \in F$ then $\pi_1 = \pi_2$. In that case, for each word $w \in \Pi(R)$, the unique π such that $\left(\begin{smallmatrix} w \\ \pi \end{smallmatrix}\right) \in F$ is called the *image* of w by F.

A class $\mathbf{L}$ of languages is said to have the *uniformisation property* if each relation $R \in \mathbf{L}$ admits a uniformisation $F \in \mathbf{L}$. We also call such a class *self-uniformisable*. The fact that **REG** (*i.e.* the class of all regular languages) has the uniformisation property is considered folklore.

3 Proof of the theorem

We now begin the proof of Theorem 2. For the rest of the section, $\mathbf{L}$ denotes a class of regular languages corresponding to a nonempty pseudo-variety $\mathbf{S}$ of semigroups, and we assume that $\mathbf{L}$ has the uniformisation property. In the following subsections, we show that $\mathbf{L}$ contains certain specific languages and allows to express more and more complex properties of words. Ultimately, we show in Subsection 3.7 that $\mathbf{L}$ can validate evaluations with respect to finite semigroups, and therefore can recognise all regular languages.

3.1 Testing letters

Recall that by Theorem 5 we know that $\mathbf{L}$ is closed under Boolean operations. Also, as $\mathbf{S}$ is nonempty, we know that every *full language* $\mathbb{A}^+$ over some alphabet $\mathbb{A}$ belongs to $\mathbf{L}$: $\mathbb{A}^+ = \alpha^{-1}(S)$, for any semigroup $S \in \mathbf{S}$, and any homomorphism from $\mathbb{A}^+$ to S.

This section is devoted to a first step of the proof: we show that $\mathbf{L}$ must be able to detect which letters appear in the given word. More formally, the main result of this section is the following lemma.

▶ **Lemma 6.** *For all alphabets $\mathbb{A}_1 \subseteq \mathbb{A}_2$, the language $\langle \mathbb{A}_1^+, \mathbb{A}_2 \rangle$ is in $\mathbf{L}$.*

One can equivalently state the above lemma by saying that $\mathbf{FO}^1[\,] \subseteq \mathbf{L}$, or that $\mathbf{S}$ contains the pseudo-variety $\mathbf{J}_1$ of finite idempotent commutative semigroups, see [5]. However, the above statement seems to better fit the rest of the presentation.

This whole subsection is devoted to a proof of Lemma 6. To prove it, notice first that it is enough to show that the semigroup $2 = \{0, 1\}$ with the operation max belongs to $\mathbf{S}$. Indeed, given two alphabets $\mathbb{A}_1 \subseteq \mathbb{A}_2$ one can consider the homomorphism α from words over $\mathbb{A}_2$ to 2, defined by $\alpha(a) = 0$ for $a \in \mathbb{A}_1$ and $\alpha(a) = 1$ otherwise. Then, $\mathbb{A}_1^+ = \alpha^{-1}(\{0\})$, and therefore belongs to $\mathbf{L}$.

Let R be the full relation between words over $\mathbb{A} = \{x\}$ and words over $\mathbb{B} = \{\square, \triangle\}$: $R = \langle (\mathbb{A} \times \mathbb{B})^+, \mathbb{A} \times \mathbb{B} \rangle$. As discussed above, R is in $\mathbf{L}$. By the assumption, $\mathbf{L}$ must contain a uniformisation F of R that is recognised by some tuple $\langle S, \alpha, T \rangle$, with $S \in \mathbf{S}$.

Let $N = N(2)$ be the number we obtain from Theorem 3 applied for S and α in the particular case $n = 2$. Consider the word $w = x^N$, and take the unique word $\pi \in \{\square, \triangle\}^N$ such that $\binom{w}{\pi} \in F$. For convenience, for all $i \in j \in N{+}1$, we write $w_{i,j}$ (resp. $\pi_{i,j}$) for the word $w(i) \dots w(j{-}1)$ (resp. $\pi(i) \dots \pi(j{-}1)$), and $s_{i,j}$ for $\alpha(\binom{w_{i,j}}{\pi_{i,j}})$.

By the definition of N, we know that there exists an idempotent e of S, and $i \in j \in k \in N{+}1$, such that $s_{i,j} = s_{j,k} = e$. Since $j - i > 0$ and $|\mathbb{B}| \geq 2$, there exists a word $\pi' \in \mathbb{B}^{j-i}$ distinct from $\pi_{i,j}$. We define $s' = \alpha(\binom{w_{i,j}}{\pi'})$.

As e is idempotent, we know that for every $\ell \geq 1$ we have $s_{0,i} \cdot e^\ell \cdot s_{k,N} = s_{0,N} \in T$. Recall that $\sharp = \sharp(S)$ is a number such that for every $s \in S$ the element $s^\sharp$ is idempotent. Consider the particular case of the above equality for $\ell = 3 \times \sharp$, we obtain:

$$\binom{w_{0,i}}{\pi_{0,i}} \cdot \left(\binom{w_{i,j}}{\pi_{i,j}} \cdot \binom{w_{i,j}}{\pi_{i,j}} \cdot \binom{w_{i,j}}{\pi_{i,j}} \right)^\sharp \cdot \binom{w_{k,N}}{\pi_{k,N}} \in F.$$

As $\pi' \neq \pi_{i,j}$ and F is a uniformisation, we know that

$$\binom{w_{0,i}}{\pi_{0,i}} \cdot \left(\binom{w_{i,j}}{\pi_{i,j}} \cdot \binom{w_{i,j}}{\pi'} \cdot \binom{w_{i,j}}{\pi_{i,j}} \right)^\sharp \cdot \binom{w_{k,N}}{\pi_{k,N}} \notin F.$$

This implies that $e' := \left(s_{i,j} \cdot s' \cdot s_{i,j}\right)^\sharp = \left(e \cdot s' \cdot e\right)^\sharp \neq e$. Now, we set $s_0 = e$ and $s_1 = e'$. As both e and e' are idempotents, we know that $s_0 \cdot s_0 = s_0$ and $s_1 \cdot s_1 = s_1$. Moreover, because e is idempotent, it is immediate to see that $s_1 \cdot s_0 = \left(e \cdot s' \cdot e\right)^\sharp \cdot e = \left(e \cdot s' \cdot e\right)^\sharp = s_1$, and that we also have $s_0 \cdot s_1 = s_1$ symmetrically. Therefore, the subset $\{s_0, s_1\}$ of S is a sub-semigroup and it is isomorphic to $\langle 2, \max\rangle$. Because $\mathbf{S}$ is stable by taking sub-semigroups and images by surjective images, $\langle 2, \max\rangle \in \mathbf{S}$. This concludes the proof of Lemma 6.

3.2 Changing alphabets

The aim of this short section is to show that $\mathbf{L}$ is strong enough not to depend on the actual alphabet of a given language. This property is expressed by the following two lemmas.

▶ **Lemma 7.** *Let $\mathbb{A}_1 \subseteq \mathbb{A}_2$ be two alphabets and $L \subseteq \mathbb{A}_1^+$. If $\langle L, \mathbb{A}_2\rangle \in \mathbf{L}$ then $\langle L, \mathbb{A}_1\rangle \in \mathbf{L}$.*

Proof. Let $\langle S, \alpha, T\rangle$ be a tuple recognising L, with $S \in \mathbf{S}$. Let βbe the homomorphism from $\mathbb{A}_1^+$ to S defined by $\beta(a) = \alpha(a)$ for $a \in \mathbb{A}_1$. Then, by the assumption, we have $L = \beta^{-1}(T)$, and therefore $\langle L, \mathbb{A}_1\rangle \in \mathbf{L}$. ◀

▶ **Lemma 8.** *Let $\mathbb{A}_1 \subseteq \mathbb{A}_2$ be two alphabets and $L \subseteq \mathbb{A}_1^+$. If $\langle L, \mathbb{A}_1\rangle \in \mathbf{L}$ then $\langle L, \mathbb{A}_2\rangle \in \mathbf{L}$.*

Proof. Let $\langle L, \mathbb{A}_1\rangle \in \mathbf{L}$ with and $\langle S_1, \alpha_1, T_1\rangle$ a tuple recognising it, with $S_1 \in \mathbf{S}$. Assume that $\mathbb{A}_1 \subseteq \mathbb{A}_2$. We prove that $\langle L, \mathbb{A}_2\rangle$ is in $\mathbf{L}$.

We showed in the proof of Lemma 6 that the semigroup $\langle 2, \max\rangle$ is in $\mathbf{S}$. This implies that $S_1 \times 2$, with the natural product operation, is also in $\mathbf{S}$. Now, let β be the homomorphism from $\mathbb{A}_2^+$ to $S_1 \times S_2$ defined by $\beta(a) = \langle \alpha_1(a), 0\rangle$ for $a \in \mathbb{A}_1$, and $\beta(a) = \langle \alpha_1(a), 1\rangle$ for $a \in \mathbb{A}_2 \setminus \mathbb{A}_1$.

It is easy to verify that $L = \beta^{-1}(T_1 \times \{0\})$, and therefore, $\langle L, \mathbb{A}_2\rangle \in \mathbf{L}$. ◀

Lemmas 7 and 8 imply that if $L \subseteq \mathbb{A}_1^+ \cap \mathbb{A}_2^+$, then $\langle L, \mathbb{A}_1\rangle \in \mathbf{L}$ iff. $\langle L, \mathbb{A}_2\rangle \in \mathbf{L}$. Therefore, we can simply say that $L \in \mathbf{L}$, without being specific about its alphabet. Thus, from that moment on we will speak simply about languages L, instead of $\langle L, \mathbb{A}\rangle$.

Using Lemma 6, we can deduce the following result:

▶ **Corollary 9.** *Let a be any letter if an alphabet $\mathbb{A}$. By $[\exists a]_\mathbb{A}$ we denote the language of words over $\mathbb{A}$ that contain at least one occurrence of a. Then $[\exists a]_\mathbb{A} \in \mathbf{L}$.*

Moreover, by $\mathbb{A}^\oplus$ if we denote the set of all words w over $\mathbb{A}$ such that each letter of $\mathbb{A}$ appears in w, then $\mathbb{A}^\oplus$ too is in $\mathbf{L}$.

Proof. It is enough to observe that $[\exists a]_\mathbb{A} = \left((\mathbb{A} \setminus \{a\})^+\right)^c$, where ^{c} denotes the complement over the full language $\mathbb{A}^+$. Lemma 6 tells us that $(\mathbb{A} \setminus \{a\})^+ \in \mathbf{L}$, and, since $\mathbf{L}$ is closed under Boolean combinations, $[\exists a]_\mathbb{A} \in \mathbf{L}$.

Now, $\mathbb{A}^\oplus = \bigcap_{a \in \mathbb{A}} [\exists a]_\mathbb{A}$, and therefore $\mathbb{A}^\oplus \in \mathbf{L}$. ◀

3.3 Counting letters

Our next step towards Theorem 2 is to notice that $\mathbf{L}$ is able to test single occurrences of letters, as expressed by the following lemma:

▶ **Lemma 10.** *Let a be a letter of an alphabet $\mathbb{A}$. Then the language of words over $\mathbb{A}$ having exactly one occurrence of a is in $\mathbf{L}$. We denote this language by $[\exists^{=1} a]_\mathbb{A}$.*

Similarly as in the case of Lemma 6, the above lemma can be equivalently expressed by saying that $\mathbf{FO}^2[\,] \subseteq \mathbf{L}$ (*i.e.* the two-variable fragment of **FO** without any predicates except equality and letter tests).

To prove the above lemma, consider three distinct letters, x, y, and z, and four distinct symbols $\otimes$, $\oplus$, $\ominus$, and $\odot$. Let R^x and R^y be the relations defined as:

$$R^x = \left\{\binom{x}{\oplus}, \binom{x}{\ominus}, \binom{y}{\otimes}, \binom{z}{\odot}\right\}^{\oplus},$$
$$R^y = \left\{\binom{x}{\otimes}, \binom{y}{\oplus}, \binom{y}{\ominus}, \binom{z}{\odot}\right\}^{\oplus}.$$

We know that R^x and R^y are in $\mathbf{L}$ because of Corollary 9. Finally, we define $R = R^x \cup R^y$, which is in $\mathbf{L}$ because it is closed under unions.

Since $\mathbf{L}$ has the uniformisation property, there exists $F \in \mathbf{L}$ uniformising R. Let $\langle S, \alpha, T\rangle$, with $S \in \mathbf{S}$, be a triple recognising F.

Now, for $p, q \in \omega$, we define L^x_p and L^y_q as the following two relations:

$$L^x_p = \left\{\binom{u}{\sigma} \in \left\{\binom{x}{\otimes}, \binom{z}{\odot}\right\}^+ \mid \otimes \text{ appears exactly } p \text{ times in } \sigma\right\},$$
$$L^y_q = \left\{\binom{v}{\tau} \in \left\{\binom{y}{\otimes}, \binom{z}{\odot}\right)\right\}^+ \mid \otimes \text{ appears exactly } q \text{ times in } \tau\right\}.$$

Notice that $\bigcup_{p\in\omega} L^x_p = \left\{\binom{x}{\otimes}, \binom{z}{\odot}\right\}^+$ and similarly $\bigcup_{q\in\omega} L^y_q = \left\{\binom{y}{\otimes}, \binom{z}{\odot}\right\}^+$.

▷ Claim 11. At least one of the two following propositions is true:

- for all $p \geq 2$ we have $\alpha(L^x_1) \cap \alpha(L^x_p) = \emptyset$,
- for all $q \geq 2$ we have $\alpha(L^y_1) \cap \alpha(L^x_q) = \emptyset$.

Proof. Assume the contrary and take:

$$p \geq 2,\ \binom{u_1}{\sigma_1} \in L^x_1, \text{ and } \binom{u_p}{\sigma_p} \in L^x_p \text{ such that } \alpha(\binom{u_1}{\sigma_1}) = \alpha(\binom{u_p}{\sigma_p}); \text{ and} \tag{1}$$
$$q \geq 2,\ \binom{v_1}{\tau_1} \in L^y_1, \text{ and } \binom{v_p}{\tau_q} \in L^y_q \text{ such that } \alpha(\binom{v_1}{\tau_1}) = \alpha(\binom{v_q}{\tau_q}). \tag{2}$$

Let w be the word $u_p \cdot v_q \cdot z$ and assume that π is the unique word over $\{\oplus, \ominus, \otimes, \odot\}$ such that $\binom{w}{\pi} \in F$. Clearly, w is in the projection of both R^x and R^y; suppose that $\binom{w}{\pi} \in R^x$ (the case $\binom{w}{\pi} \in R^y$ is symmetric). As R^x determines the symbols below the letters y and z, we know that π is of the form $\sigma \cdot \tau_q \cdot \odot$, for some word σ over $\{\oplus, \ominus, \odot\}$ of length $|u_p|$.

Consider now the new word w' over $\{x, y, z\}$ defined with $w' = u_1 \cdot v_q \cdot z$. We know that w' belongs to the projection of R^y but not to the projection of R^x, because u_1 has only one occurrence of x. Let π' be the unique word such that $\binom{w'}{\pi'} \in F$. Similarly as before, $\pi' = \sigma_1 \cdot \tau' \cdot \odot$ for some word τ' over $\{\oplus, \ominus, \odot\}$ of length $|v_q|$.

Using (1) we know that $\alpha(\binom{u_1}{\sigma_1}) = \alpha(\binom{u_p}{\sigma_p})$, and therefore $\binom{u_p}{\sigma_p} \cdot \binom{v_q}{\tau'} \cdot \binom{z}{\odot} \in F$, whose projection onto $\mathbb{A}$ equals w, which contradicts the fact that F is a uniformisation. ◀

By the symmetry, let us assume that the first item of Claim 11 holds, *i.e.* for all $\binom{u_1}{\sigma_1} \in L^x_1$ and $\binom{u_p}{\sigma_p} \in L^x_p$ with $p \geq 2$, we have $\alpha(\binom{u_1}{\sigma_1}) \neq \alpha(\binom{u_p}{\sigma_p})$.

▷ Claim 12. The language L^x_1 is in $\mathbf{L}$.

Proof. The language $L^x_0 = \{\binom{z}{\odot}\}^+$ is in $\mathbf{L}$. Therefore, $\bigcup_{p\geq 1} L^x_p = \{\binom{x}{\otimes}, \binom{z}{\odot}\}^+ \setminus L^x_0$ also belongs to $\mathbf{L}$. Thus, the above assumption about α-values implies the claim, because $L^x_1 = \alpha^{-1}(\alpha(L^x_1)) \cap \bigcup_{p\geq 1} L^x_p$. ◁

Now take $a \in \mathbb{A}$ as in the statement of Lemma 10. Consider a homomorphism β from $\mathbb{A}^+$ to $\{\binom{x}{\otimes}, \binom{z}{\odot}\}^+$ defined by $\beta(a) = \binom{x}{\otimes}$ and $\beta(b) = \binom{z}{\odot}$ for $b \neq a$. We have $[\exists^{=1} a]_{\mathbb{A}} = \beta^{-1}(F_1^x)$, and, because $\mathbf{L}$ is closed under pre-images under homomorphisms, $[\exists^{=1} a]_{\mathbb{A}}$ is in $\mathbf{L}$. This concludes the proof of Lemma 10.

With a similar – yet more technical – proof, one can show that for all $p \in \omega$, $\mathbf{L}$ contains $[\exists^{=p} a]_{\mathbb{A}}$, the language of words over $\mathbb{A}$ having exactly p letters a, but this point will not be involved in the following demonstrations. This results show that $\mathbf{L}$ must contain $\mathbf{FO}[\,]$, First-Order logic with only equalities between positions and letter tests. Recall that by Proposition 2 in [12], $\mathbf{FO}[\,]$ can be uniformised within $\mathbf{FO}[<]$. This explains why our proof of Theorem 2 needs to use the fact that $\mathbf{L}$ uniformises itself more than once.

▶ **Corollary 13.** *By modifying the homomorphism used in the proof of Lemma 10, we obtain that if $\mathbb{A}_1 \subseteq \mathbb{A}_2$ then the language $[\exists^{=1} \mathbb{A}_1]_{\mathbb{A}_2}$ of words over $\mathbb{A}_2$ that contain exactly one occurrence of a letter from $\mathbb{A}_1$ also belongs to $\mathbf{L}$.*

3.4 Order on letters

Our next goal is to introduce the order $<$ on the positions of letters in a given word. This is achieved gradually, with the first instance of the order expressed by the following lemma:

▶ **Lemma 14.** *Let $\mathbb{A}$ be an alphabet and $a_0, \dots, a_{p-1}$ be $p \geq 1$ pairwise distinct letters, that do not belong to $\mathbb{A}$. Then the language $L = \mathbb{A}^* \cdot a_0 \cdot \mathbb{A}^* \cdots \mathbb{A}^* \cdot a_{p-1} \cdot \mathbb{A}^*$ is in $\mathbf{L}$.*

Proof. The proof is quite similar to the previous one. We consider two distinct letters, x and y, and $p+1$ distinct symbols $\square, \triangle_0, \dots, \triangle_{p-1}$, and we define the relation $R := \mathbb{C}^{\oplus}$, where $\mathbb{C}$ is the alphabet $\{\binom{y}{\square}\} \cup \{\binom{x}{\triangle_i} \mid i \in p\}$.

We know that $R \in \mathbf{L}$ and therefore it admits a uniformisation $F \in \mathbf{L}$. Let $\langle S, \alpha, T\rangle$ be a triple recognising it, with $S \in \mathbf{S}$. Fix $\sharp = \sharp(S)$.

We define now the word $u = y^\sharp \cdot x \cdot y^\sharp \cdots y^\sharp \cdot x \cdot y^\sharp$, where x appears exactly p times. Since u is in the projection of R, it also belongs to the projection of F. Let π be the image of u by F, *i.e.* the unique word π satisfying $\binom{u}{\pi} \in F$. The word π is necessarily of the shape $\square^\sharp \cdot \triangle_{\sigma(0)} \cdot \square^\sharp \cdots \square^\sharp \cdot \triangle_{\sigma(p-1)} \cdot \square^\sharp$, where σ is a permutation of $p = \{0, \dots, p-1\}$.

Let $e = \alpha(\binom{y}{\square}^\sharp) \in S$. By the definition of $\sharp(S)$, e is idempotent. Consider β the homomorphism from words over $\mathbb{A}' := \mathbb{A} \sqcup \{a_i \mid i \in p\}$ to S defined by $\beta(a_i) = e \cdot \alpha(\binom{x}{\triangle_{\sigma(i)}}) \cdot e$ for $i \in p$, and $\beta(a) = e$ for $a \in \mathbb{A}$.

Now, consider σ' a second permutation of p, and w the word $w_0 \cdot a_{\sigma'(0)} \cdot w_1 \cdots w_{p-1} \cdot a_{\sigma'(p-1)} \cdot w_p \in \mathbb{A}'^*$, where the w_i's are arbitrary words over $\mathbb{A}$. Because e is idempotent, we know that

$$\beta(w) = \alpha(\tbinom{y}{\square}^\sharp \cdot \tbinom{x}{\triangle_{\sigma'(\sigma(0))}} \cdot \tbinom{y}{\square}^\sharp \cdots \tbinom{y}{\square}^\sharp \cdot \tbinom{x}{\triangle_{\sigma'(\sigma(p-1))}} \cdot \tbinom{y}{\square}^\sharp).$$

Since F is a uniformisation, $\beta(w) \in T$ if and only if σ' is the identity. Therefore,

$$\beta^{-1}(T) \cap \bigcap_{i \in p} [\exists^{=1} a_i]_{\mathbb{A}'} = \mathbb{A}^* \cdot a_0 \cdot \mathbb{A}^* \cdots \mathbb{A}^* \cdot a_{p-1} \cdot \mathbb{A}^* = L.$$

Using Lemma 10, each of the languages $[\exists^{=1}\, a_i]_{\mathbb{A}'}$ is in $\mathbf{L}$. Because $\mathbf{L}$ is closed under intersections, we can conclude that L is in $\mathbf{L}$. ◀

3.5 Subsequences

Now we need to strengthen the above lemma, to be able to compare the positions of not necessarily distinct letters. This ability is expressed by the following lemma:

▶ **Lemma 15.** *Let $\mathbb{A}$ be an alphabet, and let $a_0, \dots, a_{p-1}$ be letters of $\mathbb{A}$, with $p \geq 1$. Then the language $\mathbb{A}^* \cdot a_0 \cdot \mathbb{A}^* \cdots \mathbb{A}^* \cdot a_{p-1} \cdot \mathbb{A}^*$ is in $\mathbf{L}$. We denote this language by $[\exists a_0 < a_1 < \dots < a_{p-1}]_{\mathbb{A}}$.*

Again, this lemma is equivalent to saying that $\mathcal{B}\Sigma_1[<] \subseteq \mathbf{L}$ (*i.e.* Boolean combinations of existential First-Order sentences with the order) or equivalently that the pseudo-variety of J-trivial semigroups $\mathbf{J}$ is contained in $\mathbf{S}$.

Let $\mathbb{B} := \{\triangle_0, \dots, \triangle_{p-1}, \square\}$ be an alphabet containing $p+1$ pairwise distinct symbols. First, we consider the following relation:

$$R = \left(\begin{smallmatrix}\mathbb{A}\\ \square\end{smallmatrix}\right)^* \cdot \left(\begin{smallmatrix}a_0\\ \triangle_0\end{smallmatrix}\right) \cdot \left(\begin{smallmatrix}\mathbb{A}\\ \square\end{smallmatrix}\right)^* \cdots \left(\begin{smallmatrix}\mathbb{A}\\ \square\end{smallmatrix}\right)^* \cdot \left(\begin{smallmatrix}a_{p-1}\\ \triangle_{p-1}\end{smallmatrix}\right) \cdot \left(\begin{smallmatrix}\mathbb{A}\\ \square\end{smallmatrix}\right)^*$$

It is immediate to see that $\Pi(R)$ is exactly $[\exists a_0 < \dots < a_{p-1}]_{\mathbb{A}}$. Consider the relations $R_1 := R \cdot \left(\begin{smallmatrix}\bullet\\ \bullet\end{smallmatrix}\right) \cdot \left(\begin{smallmatrix}\mathbb{A}\\ \square\end{smallmatrix}\right)^*$ and $R_2 := \left(\begin{smallmatrix}\mathbb{A}\\ \square\end{smallmatrix}\right)^* \cdot \left(\begin{smallmatrix}\bullet\\ \bullet\end{smallmatrix}\right) \cdot R$, where $\bullet$ is a fresh letter (*i.e.* not in $\mathbb{A}$ nor in $\mathbb{B}$).

To conclude the proof of Lemma 15, we will use a fairly technical fact. It may be seen as an abstract generalisation of the technique used in the proof of Lemma 3 in [12].

▶ **Fact 16.** *Let R be a relation over a product alphabet $\mathbb{A} \times \mathbb{B}$, i.e. $R \subseteq \left(\begin{smallmatrix}\mathbb{A}\\ \mathbb{B}\end{smallmatrix}\right)^+$. Assume that $\bullet$, $\square$ are two symbols, with $\bullet \notin \mathbb{A}$. Define $R_1 := R \cdot \left(\begin{smallmatrix}\bullet\\ \bullet\end{smallmatrix}\right) \cdot \left(\begin{smallmatrix}\mathbb{A}\\ \square\end{smallmatrix}\right)^*$, $R_2 := \left(\begin{smallmatrix}\mathbb{A}\\ \square\end{smallmatrix}\right)^* \cdot \left(\begin{smallmatrix}\bullet\\ \bullet\end{smallmatrix}\right) \cdot R$, and $P := R_1 \cup R_2$. If P is in $\mathbf{L}$ then $\Pi(R)$ is in $\mathbf{L}$.*

Proof. Let $F \in \mathbf{L}$ be a uniformisation of the above relation P, and let $\langle S, \alpha, T\rangle$ be a triple recognising F, with $S \in \mathbf{S}$.

Let β be the homomorphism from $\mathbb{A}^+$ to S defined by $\beta(a) = \alpha(\left(\begin{smallmatrix}a\\ \square\end{smallmatrix}\right))$, for all $a \in \mathbb{A}$. Put $L := \Pi(R)$. Notice that if for all words w_1, w_2 in $\mathbb{A}^+$, the equality $\beta(w_1) = \beta(w_2)$ implies the equivalence $w_1 \in L \Leftrightarrow w_2 \in L$, then in fact L is in $\mathbf{L}$, because $\beta^{-1}\big(\beta(L)\big) = L \in \mathbf{L}$.

We show now that this implication holds for all w_1, w_2. Suppose that there exist $w_1 \in L$ and $w_2 \in L^{\mathrm{c}}$ such that $\beta(w_1) = \beta(w_2)$, in order to provide a contradiction.

Let w be the word $w_1 \cdot \bullet \cdot w_1$, over the alphabet $\mathbb{A} \sqcup \{\bullet\}$. This word w is in $\Pi(R_1) \cap \Pi(R_2)$, let π the unique word over $\mathbb{B} \cup \{\bullet, \square\}$ such that $\left(\begin{smallmatrix}w\\ \pi\end{smallmatrix}\right) \in F$.

We suppose for instance that $\left(\begin{smallmatrix}w\\ \pi\end{smallmatrix}\right) \in R_1$ (the case $\left(\begin{smallmatrix}w\\ \pi\end{smallmatrix}\right) \in R_2$ is symmetric). Because $\bullet \notin \mathbb{A}$, π is necessarily of the shape $\pi_1 \cdot \bullet \cdot \square^{|w_1|}$, with $\pi_1 \in \mathbb{B}^{|w_1|}$.

Let now w' be the word $w_2 \cdot \bullet \cdot w_1$ and let π' be the unique word such that $\left(\begin{smallmatrix}w'\\ \pi'\end{smallmatrix}\right) \in F$. Again, π' is of the shape $\square^{|w_2|} \cdot \bullet \cdot \pi'_1$, with $\pi'_1 \in \mathbb{B}^{|w_1|}$. Since $\beta(w_2) = \beta(w_1)$, we know that $\alpha(\left(\begin{smallmatrix}w'\\ \square^{|w_2|}\bullet\pi'_1\end{smallmatrix}\right)) = \alpha(\left(\begin{smallmatrix}w\\ \square^{|w_1|}\bullet\pi'_1\end{smallmatrix}\right))$. The latter value does not belong to T because $\left(\begin{smallmatrix}w\\ \square^{|w_1|}\bullet\pi'_1\end{smallmatrix}\right) \notin F$ – we know that F is a uniformisation and $\pi \neq \square^{|w_1|} \cdot \bullet \cdot \pi'_1$. This means that $\left(\begin{smallmatrix}w'\\ \pi'\end{smallmatrix}\right)$ is not in F, contradicting the assumption, and concluding the proof of this fact. ◀

Now we go back to the proof of Lemma 15. The letters $\left(\begin{smallmatrix}a_0\\ \triangle_0\end{smallmatrix}\right), \dots, \left(\begin{smallmatrix}a_{p-1}\\ \triangle_{p-1}\end{smallmatrix}\right), \left(\begin{smallmatrix}\bullet\\ \bullet\end{smallmatrix}\right)$ are all pairwise distinct, and none of them is in the alphabet $\mathbb{A} \times \{\square\}$. Therefore, Lemma 14 tells us that R_1 and R_2 are in $\mathbf{L}$. This means that $P := R_1 \cup R_2$ is in $\mathbf{L}$, and we can conclude with Fact 16 that $[\exists a_0 < \dots < a_{p-1}]_{\mathbb{A}} = \Pi(R)$ is in $\mathbf{L}$.

▶ **Corollary 17.** *Let $\mathbb{A}_0, \dots, \mathbb{A}_{p-1}$ be pairwise disjoint alphabets contained in $\mathbb{A}$. Then the language $L = \mathbb{A}_0^* \cdot \mathbb{A}_1^* \dots \mathbb{A}_{p-1}^* \setminus \{\epsilon\}$ is in L.*

Proof. It is enough to observe that $L = \bigcap_{i \in j \in p} \bigcap_{a_i \in \mathbb{A}_i} \bigcap_{a_j \in \mathbb{A}_j} [\exists a_j < a_i]^{\mathrm{c}}_{\mathbb{A}}$ (where $\mathbb{A}$ is the union of the $\mathbb{A}_i$'s). ◀

3.6 Polynomials

We will now prove a variant of Lemma 15 for *polynomials*. A *monomial* is a language of the shape $L_0 \cdot L_1 \cdots L_{p-1}$, where each L_i is either of the form $\mathbb{A}_i^*$, or a set of single-letter words over $\mathbb{A}_i$, and such that at least one of the L_i's is of the latter kind. An example of a monomial is the language

$$\{a,b\}^* \cdot \{x,y\} \cdot \{x\}^*.$$

Notice that the alphabets $\mathbb{A}_i$ are not required to be pairwise disjoint in that definition. A *polynomial* is a finite union of monomials. This section is devoted to a proof of the following lemma.

▶ **Lemma 18.** *Any polynomial is in* **L**.

First notice that **L** is closed under unions and therefore it is enough to prove the lemma for monomials. Consider a monomial L over an alphabet $\mathbb{A}$, *i.e.* $L = \mathbb{A}_0^{\xi_0} \cdot \mathbb{A}_1^{\xi_1} \cdots \mathbb{A}_{p-1}^{\xi_{p-1}}$, where each ξ_i is either $*$ or 1 (because $\mathbb{A}_i^1 = \mathbb{A}_i$). Take $\mathbb{A}' := \mathbb{A} \sqcup \{\bullet\}$, $\mathbb{B} := \{\Delta_0, \dots, \Delta_{p-1}, \bullet, \square\}$. Let $R := \binom{\mathbb{A}_0}{\Delta_0}^{\xi_0} \cdots \binom{\mathbb{A}_{p-1}}{\Delta_{p-1}}^{\xi_{p-1}}$, $R_1 := R \cdot \binom{\bullet}{\bullet} \cdot \binom{\mathbb{A}}{\square}^*$, and $R_2 := \binom{\mathbb{A}}{\square}^* \cdot \binom{\bullet}{\bullet} \cdot R$.

Notice that

$$R_1 = \binom{\mathbb{A}_0}{\Delta_0}^* \cdots \binom{\mathbb{A}_{p-1}}{\Delta_{p-1}}^* \cdot \binom{\bullet}{\bullet}^* \cdot \binom{\mathbb{A}}{\square}^* \cap \left[\exists^{=1} \binom{\bullet}{\bullet}\right]_{\mathbb{A}' \times \mathbb{B}} \cap \bigcap_{i \in p} \Xi_i, \tag{3}$$

where each Ξ_i is either: $\left(\mathbb{A}' \times \mathbb{B}\right)^+$ if $\xi_i = *$; or $\left[\exists^{=1} \mathbb{A}_i \times \{\Delta_i\}\right]_{\mathbb{A}' \times \mathbb{B}}$ if $\xi_i = 1$. Now, the first ingredient on the right-hand side of (3) is as in Corollary 17 and thus belongs to **L**. The other ingredients also belong to **L**, see Lemma 10, and Corollary 13. Similarly we know that $R_2 \in \mathbf{L}$. Thus, Fact 16 implies that $L = \Pi(R) \in \mathbf{L}$, which concludes the proof of Lemma 18.

▶ Remark 19. The family of polynomials is closed under union and concatenation.

3.7 Semigroups

We can now conclude the proof of Theorem 2. Let L be a regular language over some alphabet $\mathbb{A}$, that is recognised by a tuple $\langle S, \alpha, T \rangle$ with a finite semigroup S that may *a priori* not belong to **S**. Our aim is to prove that $L \in \mathbf{L}$.

Consider a word $\binom{w}{\sigma} \in \left(\mathbb{A} \times S\right)^+$ of length n. We say that such a word is an *evaluation* if for every $i \in n$ we have $\sigma(i) = \alpha\big(w(0) \cdots w(i)\big)$. Notice that in that case $w \in L$ if and only if $\sigma(n{-}1) \in T$. Let E be the set of words $\binom{w}{\sigma} \in \left(\mathbb{A} \times S\right)^+$ that are evaluations and $\sigma(n{-}1) \in T$.

▷ Claim 20. Using the above notions, we have $\Pi(E) = L$.

Proof. Every word $w \in \mathbb{A}^n$ with $n \geq 1$ admits a unique word $\sigma \in S^n$ such that $\binom{w}{\sigma}$ is an evaluation. In that case $\alpha(w) = \sigma(n{-}1)$. Thus, $w \in L$ iff. $\sigma(n{-}1) \in T$ iff. $\binom{w}{\sigma} \in E$. ◁

Our aim is to show that a variant of the set of evaluations E belongs to **L** and then invoke Fact 16 to project away the S coordinate of the evaluations.

Consider $a, b \in \mathbb{A}$ and $r, s \in S$ and define

$$I_{a,r} := \binom{a}{r} \cdot \binom{\mathbb{A}}{S}^*, \quad M_{a,r,b,s} := \binom{\mathbb{A}}{S}^* \cdot \binom{a}{r} \cdot \binom{b}{s} \cdot \binom{\mathbb{A}}{S}^*, \quad F_{a,r} := \binom{\mathbb{A}}{S}^* \cdot \binom{a}{r}.$$

Let W be the union of the languages: $I_{a,r}$ ranging over $a \in \mathbb{A}$, $r \in S$ such that $\alpha(a) \neq r$; $M_{a,r,b,s}$ ranging over those $a, b \in \mathbb{A}$, $r, s \in S$ such that $r \cdot \alpha(b) \neq s$; and $F_{a,r}$ ranging over $r \notin T$. Notice that W defined that way is a polynomial.

▷ Claim 21. The complement of W equals E.

Proof. Clearly $E \cap W = \emptyset$. Thus, it is enough to prove that if $\binom{w}{\sigma} \notin W$ then $\binom{w}{\sigma} \in E$. Let $n = |w| = |\sigma|$. Since $\binom{w}{\sigma} \notin W$, we know that $\sigma(n{-}1) \in T$ (see the languages $F_{a,r}$), thus it is enough to show that $\binom{w}{\sigma}$ is an evaluation. It is done inductively, for $i = 0, 1, \ldots, n{-}1$. The fact that $\sigma(0) = \alpha\big(w(0)\big)$ follows from the assumption that $\binom{w}{\sigma} \notin W$ (see the languages $I_{a,r}$).

Take $i < n{-}1$ and assume that $\sigma(i) = \alpha\big(w(0) \cdots w(i)\big)$. Observe that $\sigma(i{+}1)$ must equal $\sigma(i) \cdot \alpha\big(w(i{+}1)\big)$ (see the languages $M_{a,r,b,s}$). Thus, $\sigma(i{+}1) = \alpha\big(w(0) \cdots w(i{+}1)\big)$. ◁

Consider fresh letters $\bullet$ and $\bullet$, and the alphabets $\mathbb{A}' := \mathbb{A} \sqcup \{\bullet\}$, $S' := S \sqcup \{\bullet, \Delta\}$. Let:

$$R_1 := W \cdot \binom{\bullet}{\bullet} \cdot \binom{\mathbb{A}}{\Delta}^*, \qquad R_2 := \binom{\mathbb{A}}{\Delta}^* \cdot \binom{\bullet}{\bullet} \cdot W,$$
$$R_1' := R_1^{\mathrm{c}} \cap \binom{\mathbb{A}}{S}^* \cdot \binom{\bullet}{\bullet} \cdot \binom{\mathbb{A}}{\Delta}^*, \quad R_2' := R_2^{\mathrm{c}} \cap \binom{\mathbb{A}}{\Delta}^* \cdot \binom{\bullet}{\bullet} \cdot \binom{\mathbb{A}}{S}^*, \quad P := R_1' \cup R_2'.$$

Notice that both R_1 and R_2 are polynomials (see Remark 19) and therefore all the five relations defined above belong to $\mathbf{L}$.

▷ Claim 22. Using the above notions, we have

$$R_1' = E \cdot \binom{\bullet}{\bullet} \cdot \binom{\mathbb{A}}{\Delta}^*, \quad R_2' = \binom{\mathbb{A}}{\Delta}^* \cdot \binom{\bullet}{\bullet} \cdot E.$$

Proof. These equalities follow directly from the definition and Claim 21. ◁

Therefore, Fact 16 applied to P guarantees that $\Pi(E) \in \mathbf{L}$. Thus, by Claim 20 we know that $L \in \mathbf{L}$. This concludes the proof of Theorem 2.

4 Conclusions

The main result of this work shows that among pseudo-varieties of semigroups, only **REG** $=$**MSO** is strong enough to have the uniformisation property over finite words. It seems that exactly the same techniques work also for infinite words and finite trees, however the technical details are more involved there. This means that, to be able to choose witnesses in a definable way, one needs to have access to the unrestricted quantification over these witnesses (*i.e. monadic* quantifiers).

The actual arguments used in the presented proof are rather direct: they boil down to finding an appropriate relation R that is definable in the considered formalism, such that any uniformisation of R must provide some added expressive power. However, the difficulty of that reasoning lies in the deliberate design of the relations R. From this perspective, the proof can be read as a collection of instances showing how uniformisability (or generally ability to choose witnesses) leads to an increased expressive power.

The results of this work are in a sense negative, showing that all formalisms below **MSO** do not admit uniformisation. However, still some relations can be uniformised within the limited expressive power. This leads to the following decision problem:

▶ **Problem 23.** *Given a regular language R over a product alphabet $\mathbb{A} \times \mathbb{B}$, decide if R admits an* **FO**[<]*-definable uniformisation.*

Note that even if R itself is not **FO**[<]-definable, it might be the case that there is an **FO**[<]-definable uniformisation of R. The status of this problem is open at the moment.

References

1 Mikołaj Bojańczyk. Effective characterizations of tree logics. In *PODS*, pages 53–66, 2008.
2 Julius Richard Büchi. Weak Second-Order Arithmetic and Finite Automata. *Mathematical Logic Quarterly*, 6(1–6):66–92, 1960.
3 Julius Richard Büchi and Lawrence H. Landweber. Solving Sequential Conditions by Finite-State Strategies. *Transactions of the American Mathematical Society*, 138:295–311, 1969.
4 Arnaud Carayol and Christof Löding. MSO on the infinite binary tree: Choice and order. In *CSL*, pages 161–176, 2007.
5 Volker Diekert, Paul Gastin, and Manfred Kufleitner. A Survey on Small Fragments of First-Order Logic over Finite Words. *Int. J. Found. Comput. Sci.*, 19(3):513–548, 2008. doi:10.1142/S0129054108005802.
6 Samuel Eilenberg. *Automata, languages, and machines*. Pure and Applied Mathematics. Elsevier Science, 1974.
7 Calvin C. Elgot. Decision Problems of Finite Automata Design and Related Arithmetics. *Transactions of the American Mathematical Society*, 98(1):21–51, 1961.
8 Yuri Gurevich and Saharon Shelah. Rabin's uniformization problem. *Journal of Symbolic Logic*, 48(4):1105–1119, 1983.
9 Alexander Kechris. *Classical descriptive set theory*. Springer-Verlag, New York, 1995.
10 Shmuel Lifsches and Saharon Shelah. Uniformization and Skolem Functions in the Class of Trees. *Journal of Symbolic Logic*, 63(1):103–127, 1998.
11 Robert McNaughton and Seymour Papert. *Counter-free automata*. M.I.T. Press research monographs. M.I.T. Press, 1971.
12 Vincent Michielini. Uniformization Problem for Variants of First Order Logic over Finite Words. In *DLT*, pages 516–528, 2018.
13 Dominique Perrin and Jean-Éric Pin. *Infinite Words: Automata, Semigroups, Logic and Games*. Elsevier, 2004.
14 Jean-Éric Pin and Howard Straubing. Some results on C-varieties. *ITA*, 39(1):239–262, 2005. doi:10.1051/ita:2005014.
15 Michael Oser Rabin. Decidability of second-order theories and automata on infinite trees. *Transactions of the American Mathematical Society*, 141:1–35, 1969.
16 Michael Oser Rabin and Dana Scott. Finite Automata and Their Decision Problems. *IBM Journal of Research and Development*, 3(2):114–125, April 1959.
17 Alexander Rabinovich. On decidability of monadic logic of order over the naturals extended by monadic predicates. *Information and Computation*, 205(6):870–889, 2007.
18 Jan Reiterman. The Birkhoff theorem for finite algebras. *Algebra Universalis*, 14(1):1–10, 1982.
19 Marcel Paul Schützenberger. On Finite Monoids Having Only Trivial Subgroups. *Information and Control*, 8(2):190–194, 1965.
20 Dirk Siefkes. The recursive sets in certain monadic second order fragments of arithmetic. *Arch. Math. Logik*, 17(1–2):71–80, 1975.
21 Imre Simon. Piecewise testable events. In *Automata Theory and Formal Languages*, pages 214–222, 1975.
22 Imre Simon. Word Ramsey theorems. In B. Bollobas, editor, *Graph Theory and Combinatorics*, pages 283–291. Academic Press, London, 1984.
23 Denis Thérien and Thomas Wilke. Over Words, Two Variables Are as Powerful as One Quantifier Alternation. In *STOC*, pages 234–240, 1998.
24 Wolfgang Thomas. Star-Free Regular Sets of omega-Sequences. *Information and Control*, 42(2):148–156, 1979.
25 Boris A. Trakhtenbrot. Finite automata and the monadic predicate calculus. *Siberian Mathematical Journal*, 3(1):103–131, 1962.