Reversible Regular Languages: Logical and Algebraic Characterisations

Paul Gastin, Amaldev Manuel, R Govind

Reversible Regular Languages: Logical and Algebraic Characterisations

Paul Gastin
Université Paris-Saclay, ENS Paris-Saclay, CNRS, LMF, 91190, Gif-sur-Yvette, France
CNRS, ReLaX, IRL 2000, Siruseri, India

Amaldev Manuel
Indian Institute of Technology Goa, India

R. Govind
Chennai Mathematical Institute, Chennai, India
LaBRI, University of Bordeaux, France

Abstract

We present first-order (FO) and monadic second-order (MSO) logics with predicates 'between' and 'neighbour' that characterise the class of regular languages that are closed under the reverse operation and its subclasses. The ternary between predicate $\mathsf{bet}(x, y, z)$ is true if the position y is strictly between the positions x and z. The binary neighbour predicate $\mathsf{N}(x, y)$ is true when the the positions x and y are adjacent. It is shown that the class of reversible regular languages is precisely the class definable in the logics MSO(bet) and MSO(N). Moreover the class is definable by their existential fragments EMSO(bet) and EMSO(N), yielding a normal form for MSO formulas. In the first-order case, the logic FO(bet) corresponds precisely to the class of reversible languages definable in FO($<$). Every formula in FO(bet) is equivalent to one that uses at most 3 variables. However the logic FO(N) defines only a strict subset of reversible languages definable in FO($+1$). A language-theoretic characterisation of the class of languages definable in FO(N), called locally-reversible threshold-testable (LRTT), is given. In the second part of the paper we show that the standard connections that exist between MSO and FO logics with order and successor predicates and varieties of finite semigroups extend to the new setting with the semigroups extended with an involution operation on its elements. The case is different for FO(N) where we show that one needs an additional equation that uses the involution operator to characterise the class. While the general problem of characterising FO(N) is open, an equational characterisation is shown for the case of neutral letter languages.

2012 ACM Subject Classification Theory of computation → Formal languages and automata theory; Theory of computation → Logic

Keywords and phrases Regular languages, reversible languages, first-order logic, automata, semigroups

1 Introduction

In this work we look closely at the class of regular languages that are closed under the reverse operation. We fix a finite alphabet A for the rest of our discussion. The set A^* (respectively A^+) denotes the set of all (resp. non-empty) finite words over the alphabet A. If $w = a_1 \cdots a_k$ with $a_i \in A$ is a word then $w^r = a_k \cdots a_1$ denotes the reverse of w. This notion is extended to sets of words pointwise, i.e., $L^r = \{w^r \mid w \in L\}$ and we can talk about reverse of languages. A regular language $L \subseteq A^*$ is *closed under reverse* or simply *reversible* if $L^r = L$. We let Rev denote the class of all reversible regular languages. Clearly Rev is a strict subset of the class of all regular languages.

One way to look at a reversible language is as a collection of *undirected words*. When seen as first-order structures, words are directed graphs with directed edges that constitute a linear ordering on positions. If we forgo the direction then the resulting undirected graph can be read either way and hence will correspond to both the word and its reverse. Hence a set of undirected words can be equated with a reversible language and by extension the class of undirected languages can be equated with Rev.

The class Rev is easily verified to be closed under union, intersection and complementation. It is also closed under homomorphic images, and inverse homomorphic images under alphabetic (i.e., length preserving) morphisms. However it is not closed under quotients of the form $a^{-1}L := \{v \mid av \in L\}$, where a is a letter and L is a reversible language over A. For instance, the language $L = (abc)^* + (cba)^*$ is closed under reverse but the quotient $a^{-1}L = bc(abc)^*$ is not closed under reverse. Thus the class Rev fails to be a *variety* of languages — i.e., a class closed under Boolean operations, inverse homomorphic images and quotients. However reversible languages are closed under bidirectional quotients, i.e., quotients of the form $u^{-1}Lv^{-1} \cup (v^r)^{-1} L (u^r)^{-1}$, given words u, v. Thus, to a good extent, Rev shares properties similar to that of regular languages. Hence it makes sense to ask the question

> *"are there good logical characterisations for the class* Rev *and its well behaved subclasses?".*

Our results. We suggest a positive answer to the above question. We introduce two predicates *between* ($\mathsf{bet}(x, y, z)$ is true if the position y is strictly between the positions x and z) and *neighbour* ($\mathsf{N}(x, y)$ is true if the positions x and y are adjacent). The predicates *between* and *neighbour* are the natural analogues of the order relation $<$ and successor relation $+1$ in the undirected case. In fact this analogy extends to the case of logical definability. We show that Rev is the class of monadic second order (MSO) definable languages using either of the predicates, i.e., MSO(bet) or MSO(N). This is analogous to the classical Büchi-Elgot-Trakhtenbrot theorem relating regular languages and the MSO logic. Moreover, as in the Büchi-Elgot-Trakhtenbrot theorem Rev is definable in the existential MSO logics EMSO(bet) and EMSO(N).

The above analogy extends to the case of first order logic as well. We show that FO(bet) definable languages are precisely the reversible languages definable in FO($<$). Also, every formula in FO(bet) is equivalent to one that uses at most 3 variables.

However the case of FO with the neighbour relation is different. It turns out that the class of FO(N) definable languages is a strict subset of those reversible languages definable in FO($+1$). The precise characterisation of this class is one of our main contributions. A classical result on FO($+1$)-definable languages [1] states that a language is FO($+1$) definable if and only if it is a union of classes of an equivalence relation $\approx_k^t$ for some $k, t \in \mathbb{N}$, whereby two words are $\approx_k^t$-equivalent if they have identical prefixes and suffixes of length $k - 1$ and have the same subwords of length k upto threshold t (see Definition 7). For characterising FO(N)-definable languages one needs an equivalence coarser than $\approx_k^t$. We say two words are $\overset{r}{\approx}_k^t$-equivalent if they have the same prefixes and suffixes upto reverse and have the same subwords of length k upto reverse and upto threshold t (see Definition 13). It is shown that a language is definable in FO(N) if and only if it is a union of equivalence classes of $\overset{r}{\approx}_k^t$ for some $k, t \in \mathbb{N}$.

The immediate question that arises from the above characterisations is one of definability in a logic: *Given a reversible language is it definable in the logic?"*. The case of FO(bet) is decidable due to Schützenberger-McNaughton-Papert theorem that states that syntactic monoids of FO($<$) definable languages are aperiodic (equivalent to the condition that the monoid contains no groups as subsemigroups) [2, 3]. However in the case of FO(N) one needs to consider additional restrictions on the syntactic semigroups apart from those needed to characterise $\mathsf{FO}(+1)$. This is done by means of an additional involution operation (an involution $\star$ is a unary operation satisfying the laws $a^{\star\star} = a$ and $(ab)^\star = b^\star a^\star$). It is shown that syntactic semigroups of languages definable in $\mathsf{FO}(\mathsf{N})$ satisfies the equation $exe^\star = ex^\star e^\star$ where x, e are elements the semigroup and e is furthermore an idempotent. The converse direction is open in the general case. But we prove it in the restricted case of neutral letter languages. It is to be noted that the characterisation of $\mathsf{FO}(+1)$ is a tedious one that goes via categories [4].

Related work. A different but related *between* predicate (namely $a(x, y)$, for $a \in A$, is true if there is an a-labelled position between positions x and y) was introduced and studied in [5, 6, 7]. Such a predicate is not definable in $\mathsf{FO}^2(<)$, the two variable fragment of first-order logic (which corresponds to the well known semigroup variety DA [8]). The authors of [5, 6, 7] study the expressive power of $\mathsf{FO}^2(<)$ enriched with the between predicates $a(x, y)$ for $a \in A$, and show an algebraic characterisation of the resulting family of languages. The between predicate (predicates rather) in [5] is strictly less expressive than the between predicate introduced in this paper. However the logics considered in [5] have the between predicates in conjunction with order predicates $<$ and $+1$. Hence their results are orthogonal to ours.

Another line of work that has close parallels with the one in this paper is the variety theory of involution semigroups (also called $\star$-semigroups) (see [9] for a survey). Most investigations along these lines have been on subvarieties of *regular* $\star$-semigroups (i.e., $\star$-semigroups satisfying the equation $xx^\star x = x$). As far as we are aware the equation introduced in this paper has not been studied before.

Structure of the paper. In Section 2 we introduce the predicates and present our logical characterisations. This is followed by a characterisation of $\mathsf{FO}(\mathsf{N})$. In Section 3 we discuss semigroups with involution, a natural notion of syntactic semigroups for reversible languages. In Section 4 we conclude.

An extended abstract of this work appeared in [10].

2 Logics with *Between* and *Neighbour*

As usual we represent a word $w = a_1 \cdots a_n$ as a structure containing positions $\{1, \ldots, n\}$, and unary predicates P_a for each letter a in the alphabet. The predicate P_a is precisely true at those positions labelled by letter a. The atomic predicate $x < y$ (resp. $x + 1 = y$) is true if position y is after (resp. immediately after) position x. The logic FO is the logic containing atomic predicates, boolean combinations ($\phi \vee \psi$, $\phi \wedge \psi$, $\neg\psi$ whenever ϕ, ψ are formulas of the logic), and first order quantifications ($\exists x\, \psi$, $\forall x\, \psi$ if ψ is a formula of the logic). The logic MSO in addition contains second order quantification as well ($\exists X\, \psi$, $\forall X\, \psi$ if ψ is a formula of the logic) — i.e., quantification over sets of positions. By $\mathsf{FO}(\tau)$ or $\mathsf{MSO}(\tau)$ we mean the corresponding logic with atomic predicates τ in addition to the unary predicates P_a. The classical result relating MSO and regular languages states that $\mathsf{MSO}(<) = \mathsf{MSO}(+1)$ (in terms of expressiveness) defines all regular languages. We introduce two analogous predicates for the class Rev of reversible regular languages.

2.1 $\mathsf{MSO}(\mathsf{bet}), \mathsf{MSO}(\mathsf{N})$ and $\mathsf{FO}(\mathsf{bet})$

The ternary *between* predicate $\mathsf{bet}(x,y,z)$ is true for positions x,y,z when y is strictly in between x and z, i.e.,

$$\mathsf{bet}(x,y,z) \quad := \quad x < y < z \text{ or } z < y < x.$$

▶ **Example 1.** The set of all words containing $a_1a_2\cdots a_k$ or $a_ka_{k-1}\cdots a_1$ as subword is defined by the formula

$$\exists x_1 \exists x_2 \cdots \exists x_k \bigwedge_{i=1}^{k} P_{a_i}(x_i) \;\wedge\; \bigwedge_{i=2}^{k-1} \mathsf{bet}(x_{i-1}, x_i, x_{i+1}).$$

The 'successor' relation of bet is the binary predicate *neighbour* $\mathsf{N}(x,y)$ that holds true when x and y are neighbours, i.e.

$$\mathsf{N}(x,y) \quad := \quad x+1=y \text{ or } y+1=x.$$

▶ **Example 2.** A position in a word is an endpoint if it has exactly one neighbour. The following formula defines endpoints.

$$\varphi(x) := \forall y \forall z\, (\mathsf{N}(x,y) \wedge \mathsf{N}(x,z) \rightarrow y = z)$$

The set of words of even length is defined by the formula

$$\exists e_1 e_2\, \exists X (\varphi(e_1) \wedge \varphi(e_2) \wedge X(e_1) \wedge \neg X(e_2) \wedge \forall x \forall y (\mathsf{N}(x,y) \rightarrow (X(x) \leftrightarrow \neg X(y))))\ .$$

The relation $\mathsf{N}(x,y)$ can be defined in terms of bet using first-order quantifiers as $x \neq y \wedge \forall z\, \neg\mathsf{bet}(x,z,y)$. One can also define $\mathsf{bet}(x,y,z)$ in terms of N, but using second-order set quantification. To do this we assert that x,y,z are distinct positions and any subset X of positions

- that contains x, z and at least some other position
- and such that any position in X, except for x and z, has exactly two neighbours in X,

contains the position y.

▶ **Proposition 3.** *For definable languages,* $\mathsf{MSO}(\mathsf{bet}) = \mathsf{MSO}(\mathsf{N}) = \mathsf{Rev}$.

Proof. Clearly from the discussion above, $\mathsf{MSO}(\mathsf{bet}) = \mathsf{MSO}(\mathsf{N}) \subseteq \mathsf{Rev}$. To show the other inclusion, let L be a reversible regular language and let φ be a formula in $\mathsf{MSO}(<)$ defining it. Pick an endpoint e of the given word; an endpoint is a position with exactly one neighbour, a property expressible in $\mathsf{FO}(\mathsf{N}) \subseteq \mathsf{FO}(\mathsf{bet})$. We relativize the formula φ with respect to e by replacing all occurrences of $x < y$ in the formula by $(e = x \neq y) \vee \mathsf{bet}(e,x,y)$. Let $\varphi'(e)$ be the formula obtained in this way and let $\psi(e) = \neg\exists x, y\, (x \neq y \wedge \mathsf{N}(e,x) \wedge \mathsf{N}(e,y))$ be the $\mathsf{FO}(\mathsf{N})$ formula asserting that e is an endpoint, then we claim that

$$\chi = \exists e\, (\psi\,(e) \wedge \varphi'\,(e))$$

defines the language L. Let w be a word of length $k \geq 1$ then,

$$\begin{aligned} w \models \chi \;&\Leftrightarrow\; w, 1 \models \varphi'(e) \text{ or } w, k \models \varphi'(e) \\ &\Leftrightarrow\; w \models \varphi \text{ or } w^r \models \varphi \\ &\Leftrightarrow\; w \models \varphi \quad \text{(since } L \text{ is reversible).} \end{aligned}$$

Hence $L(\chi) = L(\varphi) = L$. ◀

An $\mathsf{MSO}(\tau)$ formula is in the *existential* MSO *fragment*, denoted as $\mathsf{EMSO}(\tau)$, if it is of the form $\exists X_1 \cdots \exists X_n \, \varphi$ where φ is a first-order formula over τ. In the case of words every $\mathsf{MSO}(<)$ as well as $\mathsf{MSO}(+1)$ formula is equivalent to a formula in EMSO. This extends to the case of $\mathsf{EMSO}(\mathsf{bet})$ and $\mathsf{EMSO}(\mathsf{N})$ as well.

▶ **Proposition 4.** $\mathsf{Rev} = \mathsf{EMSO}(\mathsf{bet}) = \mathsf{EMSO}(\mathsf{N})$.

Proof. Because of Proposition 3 it suffices to show that $\mathsf{Rev} \subseteq \mathsf{EMSO}(\mathsf{bet})$ and $\mathsf{Rev} \subseteq \mathsf{EMSO}(\mathsf{N})$ in terms of languages accepted.

($\mathsf{Rev} \subseteq \mathsf{EMSO}(\mathsf{bet})$)

We modify the proof of Proposition 3. We observe that in the proof the formula φ can be assumed to be in $\mathsf{EMSO}(<)$. Therefore the formula $\varphi'(e)$ is in $\mathsf{EMSO}(\mathsf{bet})$. Let us assume $\varphi'(e) = \exists X_1 \cdots \exists X_n \, \varphi''(e)$ then

$$\chi = \exists e \, (\psi(e) \wedge \exists X_1 \cdots \exists X_n \, \varphi''(e)) \equiv \exists X_1 \cdots \exists X_n \exists e \, (\psi(e) \wedge \varphi''(e)) \; .$$

Hence $\chi \in \mathsf{EMSO}(\mathsf{bet})$ is a formula accepting the language L.

($\mathsf{Rev} \subseteq \mathsf{EMSO}(\mathsf{N})$)

Let L be a language in Rev and let $\chi = \exists X_1 \cdots \exists X_n \, \varphi$ be a formula in $\mathsf{EMSO}(+1)$ defining L such that $\varphi \in \mathsf{FO}(+1)$. Let $\psi(e)$ be a formula in $\mathsf{FO}(\mathsf{N})$ that expresses the following properties:

- Every position in the word is labelled with exactly one element from the set $\{0, 1, 2\}$ indicated by the monadic predicates Y_0, Y_1, Y_2.
- Position e is an endpoint that is labelled by 0 and its neighbour is labelled by 1.
- Let x, y, z be any three consecutive positions in the word such that x and z are the neighbours of y. Then x, y, z are labelled by $i, (i{+}1) \bmod 3, (i{+}2) \bmod 3$ or $(i{+}2) \bmod 3, (i{+}1) \bmod 3, i$ in the respective order, for some $i \in \{0, 1, 2\}$.

Let χ' be the formula

$$\chi' = \exists Y_0 \exists Y_1 \exists Y_2 \exists X_1 \cdots \exists X_n \, (\varphi' \wedge \exists e \, \psi(e))$$

where φ' is obtained by replacing each occurrence of $x + 1 = y$ by the formula

$$\sigma(x, y) = \mathsf{N}(x, y) \wedge \bigvee_{i \in \{0,1,2\}} Y_i(x) \wedge Y_{(i+1) \bmod 3}(y).$$

We claim that L is recognised by χ'. Clearly if $w \models \chi$ then

$$w, Y_0 = \{1, 4, \cdots\}, Y_1 = \{2, 5, \cdots\}, Y_2 = \{3, 6, \cdots\}, e = 1 \models \psi(e) \wedge \exists X_1 \cdots \exists X_n \, \varphi' \, .$$

Hence $w \models \chi'$.

Next we claim that if $w \models \chi'$ then $w \models \chi$. Assume $w \models \chi'$ and it has length n. The only interpretations for the predicates Y_0, Y_1, Y_2 that satisfy $\psi(e)$ are either $\{1, 4, \cdots\}$, $\{2, 5, \cdots\}$, $\{3, 6, \cdots\}$ (when $e = 1$) or $\{n, n-3, \cdots\}, \{n-1, n-4, \cdots\}, \{n-2, n-5, \cdots\}$ (when $e = n$). We have two cases. When e is taken to be 1 then $x + 1 = y$ if and only if $\sigma(x, y)$ is true, and hence $w \models \chi$. When e is taken to be n, then $\sigma(x, y)$ is true if and only if $y + 1 = x$ is true. This implies that $w \models \chi''$ where χ'' is the formula obtained from χ by replacing all atomic formulas of the form $x + 1 = y$ by $y + 1 = x$. It is easy to show by induction on the structure of the formula that $w \models \chi''$ if and only if $w^r \models \chi$. Since L is closed under reverse, we deduce that $w \models \chi$. Hence the claim is proved. ◀

Proposition 3 says that $\mathsf{MSO}(\mathsf{bet}) = \mathsf{MSO}(<) \cap \mathsf{Rev}$. This carries down to the first-order case using the same relativization idea. In fact the result holds for the prefix class Σ_i (first-order formulas in prenex normal form with i blocks of alternating quantifiers starting with a $\exists$-block).

▶ **Proposition 5.** *The following is true for definable languages.*
1. $\mathsf{FO}(\mathsf{bet}) = \mathsf{FO}(<) \cap \mathsf{Rev}$.
2. $\Sigma_i(\mathsf{bet}) = \Sigma_i(<) \cap \mathsf{Rev}$.

Proof. Given an $\mathsf{FO}(<)$ formula in prenex form defining a language in Rev, we replace every occurrence of $x < y$ by $(e = x \neq y) \vee \mathsf{bet}(e, x, y)$ as before, where e is asserted to be an endpoint with $\psi(e) = \forall x, y\, \neg\mathsf{bet}(x, e, y)$. For every formula in $\Sigma_i(<)$, $i \geq 2$ this results in an equivalent formula in $\Sigma_i(\mathsf{bet})$. For the case of Σ_1, let us note that every formula in $\Sigma_1(<)$ defines a union of languages of the form $A^*a_1A^*a_2A^* \cdots A^*a_kA^*$. Such a language can be written as a disjunction of formulas like the one in Example 1. ◀

We noted that $\mathsf{FO}(\mathsf{bet}) = \mathsf{FO}(<) \cap \mathsf{Rev}$. Kamp [11] established that linear time temporal logic LTL(X, U) has the same expressive power as $\mathsf{FO}(<)$. This is used below to establish that $\mathsf{FO}(\mathsf{bet})$ has the three variable property. Formulas in LTL are built from atomic propositions using boolean connectives and the two modalities *next* (X) and *until* (U). Each formula $\varphi \in$ LTL(X, U) has an implicit free variable and is evaluated with respect to a word $w \in A^+$ and a position i in w, we write $w, i \models \varphi$ when w at position i satisfies φ. $\mathsf{X}\,\varphi$ means that φ holds at the next position, and $\varphi_1 \,\mathsf{U}\, \varphi_2$ means that φ_2 holds at some future position and φ_1 holds between the current position and this future position.

▶ **Proposition 6.** $\mathsf{FO}(\mathsf{bet}) = \mathsf{FO}^3(\mathsf{bet})$, *i.e., for each sentence in* $\mathsf{FO}(\mathsf{bet})$, *there is an equivalent sentence in* $\mathsf{FO}^3(\mathsf{bet})$ *using at most three variable names.*

Proof. It suffices to show that for every reversible language $L \subseteq A^+$ definable in LTL(X, U) there is a corresponding $\mathsf{FO}^3(\mathsf{bet})$ formula defining L.

In the proof below, we use the following macros, definable in $\mathsf{FO}^3(\mathsf{bet})$:

$$\mathsf{E}(z) = \neg\exists x, y\ \mathsf{bet}(x, z, y)$$

$$\mathsf{N}(x, y) = \neg(x = y) \wedge \neg\exists z\ \mathsf{bet}(x, z, y)$$

For each formula $\varphi \in$ LTL(X, U), we construct inductively an $\mathsf{FO}^3(\mathsf{bet})$ formula $\overline{\varphi}(x, y)$ with (at most) two free variables x and y such that for all words $w \in A^+$ and position $1 \leq i \leq |w|$ we have

$$w, i \models \varphi \qquad \text{iff} \qquad w, x \mapsto 1, y \mapsto i \models \overline{\varphi}(x, y) \tag{1}$$

The base case is when $\varphi = a \in A$ and we let $\overline{a} = P_a(y)$. For boolean connective, we define

$$\overline{\neg\varphi}(x, y) = \neg\overline{\varphi}(x, y) \qquad\qquad \overline{\varphi_1 \vee \varphi_2}(x, y) = \overline{\varphi_1}(x, y) \vee \overline{\varphi_2}(x, y)\,.$$

The interesting cases are when the top connective of the formula is a modality X or U. We give the translation for the strict version of until, defined as $\varphi_1 \,\mathsf{SU}\, \varphi_2 = \mathsf{X}(\varphi_1 \,\mathsf{U}\, \varphi_2)$. This is sufficient since we have $\mathsf{X}\,\varphi = \bot \,\mathsf{SU}\, \varphi$ and $\varphi_1 \,\mathsf{U}\, \varphi_2 = \varphi_2 \vee (\varphi_1 \wedge (\varphi_1 \,\mathsf{SU}\, \varphi_2))$. We set

$$\overline{\varphi_1 \,\mathsf{SU}\, \varphi_2}(x, y) = \exists z\Big(\big(\mathsf{bet}(x, y, z) \vee x = y\big) \wedge \overline{\varphi_2}(x, z) \wedge \forall x$$
$$\mathsf{bet}(y, x, z) \implies \exists y\big(\mathsf{bet}(y, x, z) \wedge \mathsf{E}(y) \wedge \overline{\varphi_1}(y, x)\big)\Big)$$

We should prove that (1) holds. We apply induction on the structure of the formula φ. The base case is when φ is $a \in A$. Then $w, i \models a$ iff $w, x \mapsto 1, y \mapsto i \models P_a(y)$. When φ is of the form $\neg\varphi'$ or $\varphi_1 \vee \varphi_2$ the claim (1) follows from induction hypothesis.

For the final case, let φ be $\varphi_1 \,\mathsf{SU}\, \varphi_2$. Then, $w, i \models \varphi_1 \,\mathsf{SU}\, \varphi_2$ iff there exists $k > i$ such that $w, k \models \varphi_2$ and $w, j \models \varphi_1$ for all $i < j < k$. By induction hypothesis, the latter is true if and only if $w, x \mapsto 1, z \mapsto k \models \overline{\varphi_2}(x, z)$ and $w, y \mapsto 1, x \mapsto j \models \overline{\varphi_1}(y, x)$ for all $i < j < k$. This is precisely true when $w, x \mapsto 1, y \mapsto i \models \overline{\varphi_1 \,\mathsf{SU}\, \varphi_2}(x, y)$. This finishes the proof of (1).

Let L be a reversible language defined by the formula $\varphi \in \mathrm{LTL}(\mathsf{X}, \mathsf{U})$. This means that L is reversible and $L = \{w \in A^+ \mid w, 1 \models \varphi\}$. Define the $\mathsf{FO}^3(\mathsf{bet})$ sentence

$$\Phi = \exists x (\mathsf{E}(x) \wedge \overline{\varphi}(x, x)) .$$

We claim that Φ defines the language L, which concludes the proof. Since L is the set of all words w such that $w, 1 \models \varphi$. By (1), this is precisely when $w, x \mapsto 1, y \mapsto 1 \models \overline{\varphi}(x, y)$ and by renaming when $w, x \mapsto 1 \models \overline{\varphi}(x, x)$. Since L is closed under reverse $w, x \mapsto 1 \models \overline{\varphi}(x, x)$ iff $w^r, x \mapsto 1 \models \overline{\varphi}(x, x)$ iff $w, x \mapsto n \models \overline{\varphi}(x, x)$ where n is the last position of w. Therefore L is precisely the set of all words satisfying the formula Φ. ◀

2.2 FO(N)

Next we address the expressive power of FO with the neighbour predicate.

We start by detailing the class of *locally threshold testable languages*. Recall that word y is a *factor* of word u if $u = xyz$ for some x, z in A^*. We use $\sharp(u, y)$ to denote the number of times the factor y appears in u. For $t > 0$, we define the equality with threshold t on the set $\mathbb{N}$ of natural numbers by $i =^t j$ if $i = j$ or $i, j \geq t$.

▶ **Definition 7.** *Let $\approx_k^t$, for $k, t > 0$, be the equivalence on A^*, whereby two words u and v are equivalent if either they both have length at most $k - 1$ and $u = v$, or otherwise they have*

1. *the same prefix of length $k - 1$,*
2. *the same suffix of length $k - 1$,*
3. *and the same number of occurrences, up to threshold t, for all factors of length $\leq k$, i.e., for each word $y \in A^*$ of length at most k, $\sharp(u, y) =^t \sharp(v, y)$.*

▶ **Example 8.** We have $ababab \approx_2^1 abab \not\approx_2^1 abbab$. Indeed, all the words start and end with the same letter. In the first two words the factors ab as well as ba appear at least once. While in the last word the factor bb appears once while it is not present in the word $abab$. Notice also that $ababab \not\approx_2^2 abab$ due to the factor ba.

A language is *locally threshold testable* (or LTT for short) if it is a union of $\approx_k^t$ classes, for some $k, t > 0$.

▶ **Example 9.** The language $(ab)^*$ is LTT. In fact it is *locally testable* (the special case of locally threshold testable with $t = 1$). Indeed, $(ab)^*$ is the union of three classes: $\{\varepsilon\}$, $\{ab\}$ and $abab(ab)^*$ which is precisely the set of words that begin with a, end with b, and whose only factors of length 2 are ab and ba.

A language that is definable in $\mathsf{FO}(<)$ and not LTT is $c^*ac^*bc^*$. In this language if a and b are sufficiently separated by c-blocks then the order between a and b cannot be differentiated. It can be proved that for any t, k there is a sufficiently large n such that $c^n a c^n b c^n \approx_k^t c^n b c^n a c^n$.

Locally threshold testable languages are precisely the class of languages definable in $\mathsf{FO}(+1)$ [12, 1]. Since we can define the neighbour predicate N using $+1$, clearly $\mathsf{FO}(\mathsf{N}) \subseteq \mathsf{FO}(+1) \cap \mathsf{Rev} = \mathsf{LTT} \cap \mathsf{Rev}$. But this inclusion is strict as shown in Example 11.

▶ **Example 10.** Consider the language $L = ua^* + a^*u^r$ of words that have either u as prefix and followed by an arbitrary number of a's, or u^r as suffix and preceded by an arbitrary number of a's. The language L is in $\mathsf{FO}(\mathsf{N})$. When $u = a_1 \cdots a_n$, it can be defined by a formula of the form $\exists x_1, \ldots, x_n\, \psi$ where ψ states that x_1 is an endpoint, $\bigwedge_{1 \le i < n} \mathsf{N}(x_i, x_{i+1})$, $\bigwedge_{1 < i < n} x_{i-1} \neq x_{i+1}$, $\bigwedge_{1 \le i \le n} P_{a_i}(x_i)$, and all other positions are labeled a.

▶ **Example 11.** Consider the language L over the alphabet $\{a, b, c\}$,

$$L = \{w \mid \sharp(w, ab) = 2, \sharp(w, ba) = 1 \text{ or } \sharp(w, ab) = 1, \sharp(w, ba) = 2\}.$$

Since L is locally threshold testable and reverse closed, $L \in \mathsf{FO}(+1) \cap \mathsf{Rev}$.

We can show that $L \notin \mathsf{FO}(\mathsf{N})$ by showing that the words,

$$c^k\ ab\ c^k\ ba\ c^k ab\ c^k \in L \qquad\qquad c^k\ ab\ c^k\ ab\ c^k\ ab\ c^k \notin L$$

for $k > 0$ are indistinguishable by an $\mathsf{FO}(\mathsf{N})$ formula of quantifier depth k. For showing the latter claim, one uses Ehrenfeucht-Fraissé games and argues that in the k-round EF-game the duplicator has a winning strategy. The strategy is roughly described below:

$$\underline{c^k abc^k b}\, ac^k abc^k \qquad\qquad c^k abc^k a\, \underline{bc^k abc^k}$$

Any move of the spoiler is mimicked by the duplicator in the corresponding underlined or non-underlined part of the other word, while maintaining the neighbourhood relation between positions. For instance, if the spoiler plays the first b on the underlined part of the first word, then the duplicator chooses the last b on the underlined portion of the word on the right. Similarly, if the spoiler plays the first a on the non-underlined part of the first word, the duplicator chooses the last a on the non-underlined portion of the word on the right. Note that, since no order on positions in the words can be checked with the neighbour predicate, there is no way to distinguish between these words, if the duplicator plays in the above way ensuring that the position played has the same neighbourhood relation as the position played by the spoiler. Therefore, the Neighbour predicate will not be able to distinguish between ab and ba when they are sufficiently separated by c's.

From the above example, we get,

▶ **Proposition 12.** *For definable languages,* $\mathsf{FO}(\mathsf{N}) \subsetneq \mathsf{FO}(+1) \cap \mathsf{Rev} = \mathsf{LTT} \cap \mathsf{Rev}$.

Next we will characterise the class of languages accepted by $\mathsf{FO}(\mathsf{N})$. Recall that $\sharp(w, v)$ denotes the number of occurrences of v in w, i.e., the number of pairs (x, y) such that $w = xvy$. We extend this to $\sharp^r(w, v)$ that counts the number of occurrences of v or v^r in w, i.e., the number of pairs (x, y) such that $w = xvy$ or $w = xv^r y$. Notice that $\sharp^r(w, v) = \sharp^r(w, v^r) = \sharp^r(w^r, v) = \sharp^r(w^r, v^r)$.

▶ **Definition 13.** *We define now the* locally-reversible threshold testable *(*LRTT*) equivalence relation. Let $k, t > 0$. Two words $w, w' \in A^*$ are (k, t)-*LRTT *equivalent, denoted $w \overset{r}{\approx}{}_k^t w'$ if $|w| < k$ and $w' \in \{w, w^r\}$, or*

- *w, w' are both of length at least k, and*
- *$\sharp^r(w, v) =^t \sharp^r(w', v)$ for all $v \in A^{\le k}$, and*
- *if x, x' are the prefixes of w, w' of length $k-1$ and y, y' are the suffixes of w, w' of length $k-1$ then $\{x, y^r\} = \{x', y'^r\}$.*

Notice that $w \overset{r}{\approx}^t_k w^r$ for all $w \in A^*$ and $w \approx^t_k w'$ implies $w \overset{r}{\approx}^t_k w'$ for all $w, w' \in A^*$. Notice also that $\overset{r}{\approx}^t_k$ is not a congruence. Indeed, we have $ab \overset{r}{\approx}^t_k ba$ but $aba \overset{r}{\not\approx}^t_k baa$. On the other hand, if $v \overset{r}{\approx}^t_k w$ then for all $u \in A^*$ we have $uv \overset{r}{\approx}^t_k uw$ or $uv \overset{r}{\approx}^t_k uw^r$, and similarly $vu \overset{r}{\approx}^t_k wu$ or $vu \overset{r}{\approx}^t_k w^r u$.

▶ **Definition 14** (Locally-Reversible Threshold Testable Languages). *A language L is* locally-reversible threshold testable*,* LRTT *for short, if it is a union of equivalence classes of $\overset{r}{\approx}^t_k$ for some $k, t > 0$.*

▶ **Theorem 15.** *Languages defined by* FO(N) *are precisely the class of locally-reversible threshold testable languages.*

Proof. ($\Leftarrow$) Assume we are given an LRTT language, i.e., a union of $\overset{r}{\approx}^t_k$-classes for some $k, t > 0$. We explain how to write an FO(N) formula for each $\overset{r}{\approx}^t_k$-class. Consider a word $v = a_1 a_2 \cdots a_n \in A^+$. For $m \in \mathbb{N}$, we can say that v or its reverse occurs at least m times in a word $w \in A^*$, i.e., $\sharp^r(w, v) \geq m$, by the formula

$$\begin{aligned}\varphi_v^{\geq m} = \exists x_{1,1} \cdots \exists x_{1,n} \cdots \exists x_{m,1} \cdots \exists x_{m,n} \\ \bigwedge_{i=1}^{m} \Big(\bigwedge_{j=1}^{n-1} \mathsf{N}(x_{i,j}, x_{i,j+1}) \wedge \bigwedge_{j=2}^{n-1} (x_{i,j-1} \neq x_{i,j+1}) \wedge \bigwedge_{j=1}^{n} P_{a_j}(x_{i,j}) \Big) \\ \wedge \bigwedge_{1 \leq i < j \leq m} \neg((x_{i,1} = x_{j,1} \wedge x_{i,n} = x_{j,n}) \vee (x_{i,1} = x_{j,n} \wedge x_{i,n} = x_{j,1}))\,.\end{aligned}$$

Similarly, we can write a formula $\psi_v \in$ FO(N) that says that a word belongs to $\{v, v^r\}$. Finally, given two words of same length $u, v \in A^n$, we can write a formula $\chi_{u,v} \in$ FO(N) that says that u, v occur at two different end points of a word w, i.e., that $\{x, y^r\} = \{u, v\}$ where x, y are the prefix and suffix of w of length n.

($\Rightarrow$) Hanf's theorem ([13], Theorem 2.4.1) states that two first-order structures $\mathfrak{A}$ and $\mathfrak{B}$ are m-equivalent (i.e., indistinguishable by any FO formula of quantifier rank at most m), for some $m \in \mathbb{N}$ if for each 3^m ball type S, both $\mathfrak{A}$ and $\mathfrak{B}$ have the same number of 3^m balls of type S up to a threshold $m \times e$, where $e \in \mathbb{N}$. Models of FO(N) formulas are first-order structures of the form $(\{1, \ldots, n\}, (P_a)_{a \in \Sigma}, \mathsf{N})$ that are labelled undirected path graphs. Balls in such a graph are nothing but factors of the corresponding undirected word. Applying Hanf's theorem to undirected path graphs, we obtain that given an FO(N) formula Φ, there exist $k, t > 0$ such that if two words w and w' are $\overset{r}{\approx}^t_k$-equivalent, then w satisfies Φ if and only if w' satisfies Φ. Therefore, the set of all words satisfying Φ is an LRTT language. ◀

3 The Membership problem for the Logics

In this section we address the question of definability of a language — "is the given reversible regular language definable by a formula in the logic?" — in the previously defined logics. We show that in the case of FO(bet) the existing theorems provide an algorithm for the problem, while for FO(N) the answer is not yet known.

3.1 Membership in MSO(bet), MSO(N), FO(bet)

By Proposition 3, to check if a regular language is definable in MSO(bet) or in MSO(N) it suffices to check if it is reversible. Next we look at the membership problem for FO(bet).

First we recall the notion of recognisability by a finite semigroup. A finite semigroup $(S, \cdot)$ is a finite set S with an associative binary operation $\cdot \colon S \times S \to S$. If the semigroup operation has an identity, then it is necessarily unique and is denoted by 1. In this case S is called a monoid. A semigroup morphism from $(S, \cdot)$ to $(T, +)$ is a map $h \colon S \to T$ that preserves the semigroup operation, i.e., $h(a \cdot b) = h(a) + h(b)$ for a, b in S. Further if S and T are monoids the map is a monoid morphism if h maps the identity of S to the identity of T.

The set A^+ under concatenation forms a free semigroup while A^* under concatenation forms a free monoid with the empty word ε as the identity. A language $L \subseteq A^*$ is *recognised* by a monoid $(M, \cdot)$, if there is a morphism $h \colon A^* \to M$ and a set $P \subseteq M$, such that $L = h^{-1}(P)$.

Given a language L, the *syntactic congruence* of L, denoted as $\sim_L$ is the congruence on A^*,

$$x \sim_L y \quad \text{if} \quad uxv \in L \Leftrightarrow uyv \in L \text{ for all } u, v \in A^*. \tag{2}$$

The quotient $A^*/{\sim_L}$, denoted as $M(L)$, is called the *syntactic monoid*. It recognises L and is the unique minimal object with the following canonicity property: any monoid M recognising L has a *surjective* morphism from a submonoid of M to $M(L)$ [4].

A semigroup (or monoid) is aperiodic if there is some $n \in \mathbb{N}$ such that $a^n = a^{n+1}$ for each element a of the semigroup. Schützenberger-McNaughton-Papert theorem [2, 3] states that a language L is definable in $\mathsf{FO}(<)$ if and only if the syntactic monoid of L is aperiodic. This theorem in conjunction with Proposition 5 gives that,

▶ **Corollary 16.** *A reversible language L is definable in* $\mathsf{FO}(\mathsf{bet})$ *if and only if $M(L)$ is aperiodic.*

The above result hence yields an algorithm for definability of a language in $\mathsf{FO}(\mathsf{bet})$, i.e., check if the language is reversible, if so compute the syntactic monoid and test for aperiodicity.

3.2 Membership in $\mathsf{FO}(\mathsf{N})$

Next we look at the membership problem for the logic $\mathsf{FO}(\mathsf{N})$. The corresponding problem for $\mathsf{FO}(+1)$ is known only in terms of syntactic semigroups that we recall now. A language $L \subseteq A^+$ is *recognised* by a semigroup $(S, \cdot)$, if there is a morphism $h \colon A^+ \to S$ and a set $P \subseteq S$, such that $L = h^{-1}(P)$.

The *syntactic congruence* of $L \subseteq A^+$, denoted as $\sim_L$, is the congruence on A^+ given by Equation (2). The quotient $A^+/{\sim_L}$, denoted as $S(L)$, is called the *syntactic semigroup*. It shares the canonicity property of syntactic monoids, namely it recognises L and is the unique minimal object that has a surjective morphism from a subsemigroup of any semigroup recognising L [14].

The characterisation theorem for $\mathsf{FO}(+1)$ due to Brzozowski and Simon [15], and Beauquier and Pin [12], is stated below. Recall that an element of a semigroup e is an *idempotent* if $e \cdot e = e$.

▶ **Theorem 17** (Brzozowski-Simon, Beauquier-Pin)**.** *The following are equivalent for a language $L \subseteq A^+$.*

1. *L is locally threshold testable.*
2. *L is definable in* $\mathsf{FO}(+1)$.
3. *The syntactic semigroup of L is finite, aperiodic and satisfies the identity $e\,x\,f\,y\,e\,z\,f = e\,z\,f\,y\,e\,x\,f$ for all $e, f, x, y, z \in S(L)$ with e, f idempotents.*

Because of Proposition 12 it is clear that we need to add more identities to characterise the logic FO(N).

In the particular case of reversible languages the syntactic semigroups described above admit further properties. The observation is that the reverse operation extends to congruence classes of the syntactic congruence as shown next. Fix a reversible language L. Let $[x] \in S(L)$ denote the equivalence class of a word $x \in A^+$ under the syntactic congruence. Then we let $[x]^r = [x^r]$. This is well defined since $x \sim_L y$ if and only if $x^r \sim_L y^r$. Furthermore this map admits two properties — it is an involution (a map that is its own inverse), since

$$([x]^r)^r = ([x^r])^r = [(x^r)^r] = [x]\,, \tag{3}$$

and it is an anti-automorphism on the semigroup $S(L)$ since

$$([x] \cdot [y])^r = ([x \cdot y])^r = [(x \cdot y)^r] = [y^r \cdot x^r] = [y^r] \cdot [x^r]\ . \tag{4}$$

Thus $S(L)$ is a semigroup with an involution operation, namely the reverse. Formally, a *semigroup with involution* (also called a $\star$-semigroup) $(S, \cdot, \star)$ is a semigroup $(S, \cdot)$ extended with an operation $\star\colon S \to S$ (called the involution) such that

1. the operation $\star$ is an involution on S, i.e., $(a^\star)^\star = a$ for all elements a of S,
2. the operation $\star$ is an anti-automorphism on S (isomorphism between S and opposite of S), i.e., $(a \cdot b)^\star = b^\star \cdot a^\star$ for any a, b in S.

It is a $\star$-monoid if S is a monoid. An element x in a $\star$-semigroup is called *hermitian* if it is its own involution, i.e. $x^\star = x$. It is easy to see that in the case of $\star$-monoids, necessarily $1^\star = 1$, i.e. identity is hermitian. Similarly if the semigroup has a zero it is hermitian as well.

In the light of this definition we call $S(L)$ the *syntactic $\star$-semigroup* of a reversible language L. Next we show that syntactic $\star$-semigroups of FO(N)-definable languages obey the identity $exe^\star = ex^\star e^\star$, where x is an element of the semigroup and e is an idempotent of the semigroup. Before we prove it, we look at a couple of examples.

▶ **Example 18.** Fix the alphabet $\{c, x, y\}$ for the example below. Consider the reversible languages $L_1 = c^*xyc^*xyc^* + c^*yxc^*yxc^*$, $L_2 = c^*(xy + yx)c^*(xy + yx)c^*$. It is easy to verify that both languages are definable in FO(+1). Their syntactic semigroups are shown in Figure 1. These semigroups were computed using the online tool of Charles Paperman [16].

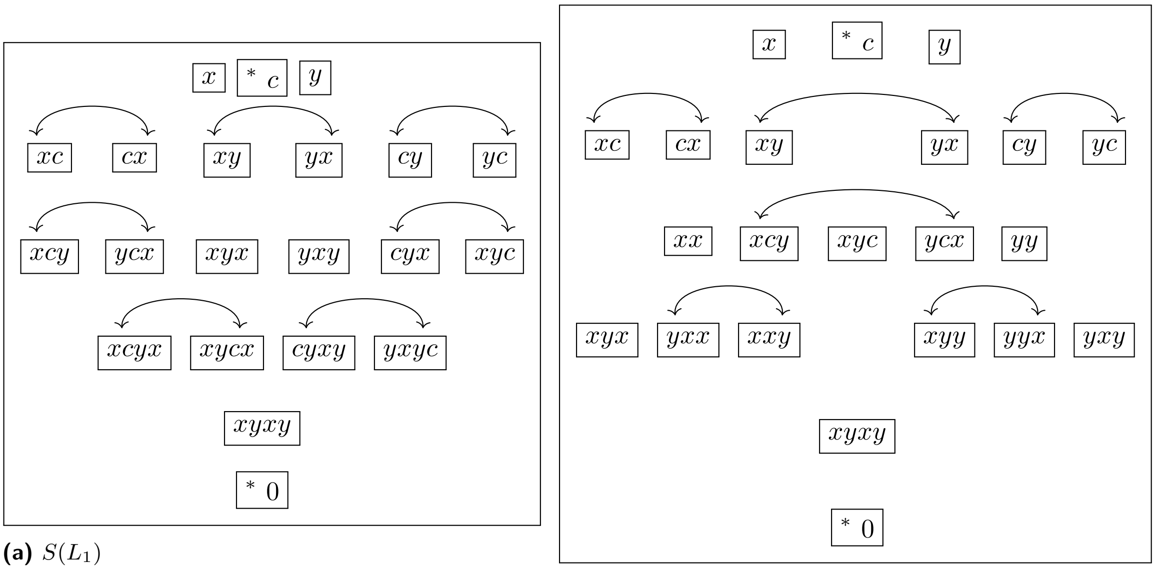
Let x, y, c also denote the images of the corresponding letters in the syntactic semigroups, which are indeed hermitian. Clearly the syntactic semigroups are generated by these elements. It is easy to deduce that c is an idempotent while x, y are not. Also both semigroups have zeros (for instance any product x^k involving more than two occurrences of x is a zero). Next we have a closer look at them.

1. We claim that $S(L_1)$ does not satisfy the identity $exe^\star = ex^\star e^\star$. Consider the two words $cxyc$ and $cyxc$. Note that $(xy)^\star = y^\star x^\star = yx$. It suffices to show that $cxyc \not\sim_{L_1} cyxc$, that is evident since $cxy \cdot cxyc \cdot \varepsilon \in L_1$ while $cxy \cdot cyxc \cdot \varepsilon \notin L_1$.
2. Next we verify that $S(L_2)$ satisfies the identity $exe^\star = ex^\star e^\star$. Since the only two idempotents are c and 0, it suffices to show that $cuc \sim_{L_2} cu^r c$ for all nonempty words u. It is easy to verify that $pcucq \in L$ if and only if $pcu^r cq \in L$ for all words p, q and hence $cuc \sim_{L_2} cu^r c$. Since u was arbitrary it follows that $S(L_2)$ satisfies the identity $exe^\star = ex^\star e^\star$.

▶ **Theorem 19.** *The syntactic $\star$-semigroup of an* FO(N)*-definable language satisfies the identity*

$$exe^\star = ex^\star e^\star,$$

where e is an idempotent, and x is any element of the semigroup.



(a) $S(L_1)$

(b) $S(L_2)$

Figure 1 Syntactic involution semigroups of $L_1 = c^*xyc^*xyc^* + c^*yxc^*yxc^*$ and $L_2 = c^*(xy + yx)c^*(xy+yx)c^*$ over the alphabet $\{c, x, y\}$. Idempotents are indicated by $*$. Involutions on elements are indicated by arrows, unless the element is hermitian.

Proof. Assume we are given an FO(N)-language L, with its syntactic $\star$-semigroup $S(L) = (A^+/{\sim_L}, \cdot, \star)$, and $h\colon A^+ \to S(L)$ the canonical morphism recognising L. Let e be an idempotent of $S(L)$, and let x be an element of $S(L)$. Pick nonempty words u and s such that $h(u) = e$ and $h(s) = x$.

By definition of the involution, $h(u^r) = e^\star$ and $h(s^r) = x^\star$. We are going to show that $usu^r \sim_L us^ru^r$ and hence they will correspond to the same element in the syntactic $\star$-semigroup, proving that $exe^\star = ex^\star e^\star$.

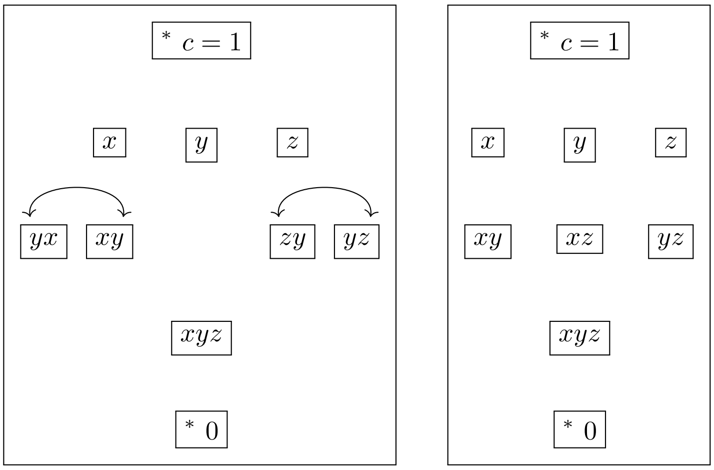
Since L is FO(N) definable, we know by Theorem 15 that L is a union of $\overset{r}{\approx}{}^t_k$ equivalence classes for some $k, t > 0$. Consider the words $w = (u^k)s(u^r)^k$ and $w^r = (u^k)s^r(u^r)^k$, obtained by pumping the words corresponding to e and $e^\star$. Since $e, e^\star$ are idempotents, it is clear that $h(w) = h(usu^r) = exe^\star$ and $h(w^r) = h(us^ru^r) = ex^\star e^\star$.

For all contexts $\alpha, \beta \in A^*$, we show below that $\alpha w\beta \overset{r}{\approx}{}^t_k \alpha w^r\beta$, which implies $\alpha w\beta \in L$ iff $\alpha w^r\beta \in L$ since L is a union of $\overset{r}{\approx}{}^t_k$ classes. It follows that $w \sim_L w^r$ and therefore $h(w) = h(w^r)$, that will conclude the proof.

Fix some contexts $\alpha, \beta \in A^*$. Since $u \neq \varepsilon$, the words $\alpha w\beta$ and $\alpha w^r\beta$ have the same prefix of length $k-1$ and the same suffix of length $k-1$. Now, consider $v \in A^k$. If an occurrence of v (resp. v^r) in $\alpha w\beta$ overlaps with α or β then we have the very same occurrence in $\alpha w^r\beta$. Using $w \overset{r}{\approx}{}^t_k w^r$, we deduce that $\sharp^r(\alpha w\beta, v) =^t \sharp^r(\alpha w^r\beta, v)$. Therefore, $\alpha w\beta \overset{r}{\approx}{}^t_k \alpha w^r\beta$. ◀

The converse direction is open. The similar direction in the case of FO(+1) goes via categories [17] and uses the Delay theorem of Straubing [18, 4]. However in the special case when the syntactic $\star$-semigroups are monoids (i.e. contains an identity) we can get an easy converse.

Let A be an alphabet and let $L \subseteq A^+$ be a language over A. A letter $c \in A$ is *neutral* in the language L if $xy \in L \Leftrightarrow xcy \in L$ for all $x, y \in A^*$ such that $|xy| \geq 1$, i.e. membership in L is invariant under insertion or deletion of the letter c. By definition, it is easy to see that if L has a neutral letter then that maps to an element that is identity in the syntactic



(a) $S(L_3)$ **(b)** $S(L_4)$

Figure 2 Syntactic involution semigroups of $L_3 = c^*xc^*yc^*zc^* + c^*zc^*yc^*xc^*$ and $L_4 = \text{Permutations}(L_3)$ over the alphabet $\{c, x, y, z\}$. Idempotents are indicated by $*$. Involutions on elements are indicated by arrows, unless the element is hermitian.

semigroup of L. For aperiodic semigroups the converse is also true.

▶ **Lemma 20.** *Let L be an aperiodic language. Then the syntactic semigroup $S(L)$ contains an identity if and only if L has a neutral letter.*

Proof. ($\Leftarrow$) By definition.
($\Rightarrow$) Assume $S(L)$ has an identity and let $cu \in A^+$, where $c \in A$, be a word mapping to it. If u is empty then c is a neutral letter and we are done. Otherwise let $a = \varphi(c)$ and $b = \varphi(u)$. Then $aab = a$ and by repeated substitution $a^n ab^n = a$ and since $S(L)$ is aperiodic there is some n such that $a^n a = a^n$ and hence $a = a^n ab^n = a^n b^n = ab$. Therefore c is a neutral letter. ◀

▶ **Example 21.** Fix the alphabet $\{c, x, y, z\}$ for the example below. Let $L_3 = c^*xc^*yc^*zc^* + c^*zc^*yc^*xc^*$, and L_4 be the set of all permutations of words in L_3, ie. the commutative closure of L_3. Both the languages are closed under reverse. Moreover it is easy to verify that both are definable in $\mathsf{FO}(<)$ and by extension in $\mathsf{FO}(\mathsf{bet})$. Hence their syntactic semigroups are aperiodic. They are shown in Figure 2.

Let x, y, z, c denote the images of the corresponding letters in the syntactic semigroups. Clearly the syntactic semigroups are generated by these elements. Since letter c is neutral in both L_3 and L_4, we deduce that c is the identity. Also any product involving at least two occurrences of x (or y, or z) is a (non-accepting) zero element denoted as 0. These are the only idempotents in the syntactic semigroups. Next we have a closer look at them.

1. Consider the language L_3. The semigroup $S(L_3)$ is a monoid with identity c. It satisfies the additional rules $xz = zx = 0$ and $xyz = zyx$. Since $cxcccyc = xy \neq yx = cycccxc$, $S(L_3)$ does not satisfy the condition on syntactic semigroups given by Theorem 17 and hence L_3 is not LTT, and by extension not in LRTT either.
2. Next consider the language L_4 that is all the permutations of words in L_1. The semigroup $S(L_4)$ is commutative and it has an identity (the element c) and a zero. Since all elements

in $S(L_4)$ are hermitian, it is clear that $S(L_4)$ satisfies the identity $exe^\star = ex^\star e^\star$ in addition to those corresponding to FO(+1).

▶ **Lemma 22.** *If all elements of an involution semigroup S are hermitian, then S is commutative. Conversely if S is commutative and generated by a subset of its hermitian elements, then all elements of S are hermitian.*

Proof. If all elements of S are hermitian, then $ab = (ab)^\star = b^\star a^\star = ba$, for all elements $a, b \in S$, i.e. S is commutative. Conversely assume S is commutative and generated by a subset of hermitian elements. Then every element can be written as a product $x_1 \cdots x_n$ where each x_i is hermitian. Then $(x_1 \cdots x_n)^\star = x_n^\star \cdots x_1^\star = x_n \cdots x_1 = x_1 \cdots x_n$. Hence all elements are hermitian. ◀

Since syntactic semigroups are generated by images of letters (that are clearly hermitian), we obtain the following.

▶ **Proposition 23.** *Let $L \subseteq A^+$ be a reversible language with a neutral letter. Then the following are equivalent.*

1. *All elements of $S(L)$ are hermitian.*
2. *$S(L)$ satisfies the identity $exe^\star = ex^\star e^\star$.*
3. *$S(L)$ is locally idempotent i.e., it satisfies the identity $exeye = eyexe$ for all idempotents e and elements $x \in S(L)$.*
4. *$S(L)$ is commutative.*

Proof. Equivalence of (1) and (4) is from Lemma 22. Now, (1) implies (2) is clear and the converse (2) implies (1) is because $S(L)$ has an identity. Similarly (4) implies (3) is clear and (3) implies (4) is due to the presence of an identity. ◀

▶ **Corollary 24.** *Let L be a reversible language with a neutral letter. The following are equivalent.*

1. *$L \in$ LTT, equivalently, L is definable in FO(+1).*
2. *$L \in$ LRTT, equivalently, L is definable in FO(N).*
3. *$L \in$ ACom (the class of aperiodic and commutative languages), equivalently, L is definable in FO(=) [4].*

Proof. ($1 \Rightarrow 3$) and ($2 \Rightarrow 3$) follows from Proposition 23. The converse inclusions are by definition. ◀

4 Conclusion

The logics MSO(bet), MSO(N) and FO(bet) behave analogously to their classical counterparts MSO(<), MSO(+1) and FO(<). But the logic FO(N) gives rise to a new class of languages, locally-reversible threshold testable languages. The quest for characterising the new class takes us to the formalism of involution semigroups. The full characterisation of the new class is the main question we leave open. It would also be interesting to know what are the natural analogues of standard fragments of FO and their expressive power, for instance classes defined by bounded number of variables, in the reversible world. Another line of investigation is to study the equationally-defined classes that arise naturally from automata theory.

References

1 Thomas W. Classifying Regular Events in Symbolic Logic. *J. Comput. Syst. Sci.*, 1982. **25**(3):360–376. doi:10.1016/0022-0000(82)90016-2. URL https://doi.org/10.1016/0022-0000(82)90016-2.

2 Schützenberger MP. On Finite Monoids Having Only Trivial Subgroups. *Information and Control*, 1965. **8**(2):190–194. doi:10.1016/S0019-9958(65)90108-7. URL https://doi.org/10.1016/S0019-9958(65)90108-7.

3 McNaughton R, Papert SA. Counter-Free Automata (M.I.T. Research Monograph No. 65). The MIT Press, 1971. ISBN 0262130769.

4 Straubing H. Finite Automata, Formal Logic, and Circuit Complexity. Birkhäuser Verlag, Basel, Switzerland, 1994. ISBN 3-7643-3719-2.

5 Krebs A, Lodaya K, Pandya PK, Straubing H. Two-variable Logic with a Between Relation. In: Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science, LICS '16, New York, NY, USA, July 5-8, 2016. 2016 pp. 106–115. doi:10.1145/2933575.2935308. URL https://doi.org/10.1145/2933575.2935308.

6 Krebs A, Lodaya K, Pandya PK, Straubing H. An Algebraic Decision Procedure for Two-Variable Logic with a Between Relation. In: 27th EACSL Annual Conference on Computer Science Logic, CSL 2018, September 4-7, 2018, Birmingham, UK. 2018 pp. 28:1–28:17. doi:10.4230/LIPIcs.CSL.2018.28. URL https://doi.org/10.4230/LIPIcs.CSL.2018.28.

7 Krebs A, Lodaya K, Pandya PK, Straubing H. Two-variable logics with some betweenness relations: Expressiveness, satisfiability and membership. *arXiv preprint arXiv:1902.05905*, 2019.

8 Tesson P, Therien D. Diamonds Are Forever: The Variety **DA**. In: Semigroups, Algorithms, Automata and Languages, Coimbra (Portugal) 2001. World Scientific, 2002 pp. 475–500.

9 Crvenković S, Dolinka I. Varieties of involution semigroups and involution semirings: a survey. In: Proceedings of the International Conference "Contemporary Developments in Mathematics" (Banja Luka, 2000), Bulletin of Society of Mathematicians of Banja Luka. 2000 pp. 7–47.

10 Gastin P, Manuel A, Govind R. Logics for Reversible Regular Languages and Semigroups with Involution. In: Hofman P, Skrzypczak M (eds.), Proceedings of the 23th International Conference on Developments in Language Theory (DLT'19), volume 11647 of *Lecture Notes in Computer Science*. Springer, Warsaw, Poland, 2019 pp. 182–191.

11 Kamp JAW. Tense Logic and the Theory of Linear Order. Ph.D. thesis, University of California, 1968.

12 Beauquier D, Pin JÉ. Languages and Scanners. *Theor. Comput. Sci.*, 1991. **84**(1):3–21. doi:10.1016/0304-3975(91)90258-4. URL https://doi.org/10.1016/0304-3975(91)90258-4.

13 Ebbinghaus HD, Flum J. Finite Model Theory: Second Edition. Springer Monographs in Mathematics. Springer Berlin Heidelberg, 2005. ISBN 9783540287889.

14 Pin JÉ. Mathematical Foundations of Automata Theory. URL https://www.irif.fr/~jep/PDF/MPRI/MPRI.pdf.

15 Brzozowski JA, Simon I. Characterizations of Locally Testable Events. In: 12th Annual Symposium on Switching and Automata Theory, East Lansing, Michigan, USA, October 13-15, 1971. 1971 pp. 166–176. doi:10.1109/SWAT.1971.6. URL https://doi.org/10.1109/SWAT.1971.6.

16 Paperman C. Semigroup Online. URL https://www.paperman.name/semigroup/.

17 Tilson B. Categories as algebra: An essential ingredient in the theory of monoids. *Journal of Pure and Applied Algebra*, 1987. **48**(1-2):83–198.

18 Straubing H. Finite semigroup varieties of the form $\mathbf{V} \star \mathbf{D}$. *Journal of Pure and Applied Algebra*, 1985. **36**:53 – 94.