Acta Informatica manuscript No.
(will be inserted by the editor)

First-order logics: some characterizations and closure properties

Christian Choffrut · **Andreas Malcher**
Carlo Mereghetti · **Beatrice Palano**

Received: date / Accepted: date

Abstract The characterization of the class of FO$[+]$-definable languages by some generating or recognizing device is still an open problem. We prove that, restricted to word bounded languages, this class coincides with the class of semilinear languages. We also study the closure properties of the classes of languages definable in FO$[+1]$, FO$[<]$, FO$[+]$ and FOC$[+]$ under the main classical operations.

Keywords Word bounded languages · Semilinear sets · First-order logic · Counting quantifier · Shuffle · Concatenation

1 Introduction

The aim of descriptive complexity is to provide logical characterizations of relevant classes of languages. The first result in this area dates back to Büchi [4] who gave a characterization of regular languages via monadic second order logic. Since then, a well consolidated trend in the literature provides characterizations of several language classes via different logics. This has also other consequences: a logical description of a language often leads to a precise estimate of the parallel complexity of membership and related problems for that language (see, e.g., [1,18]).

This work was partially supported by the Italian MURST under the project "PRIN: Aspetti matematici e applicazioni emergenti degli automi e dei linguaggi formali: metodi probabilistici e combinatori in ambito di linguaggi formali", and by CRUI/DAAD under the project "Programma Vigoni: Reducing complexity by introducing structure".
A preliminary version has been published in the *Proceedings of the 4th International Conference on Language and Automata Theory and Applications (LATA 2010).*

Christian Choffrut
LIAFA, UMR 7089, 175 Rue du Chevaleret, Paris 13, France
E-mail: fcc@liafa.jussieu.fr

Andreas Malcher
Institut für Informatik, Universität Giessen, Arndtstr. 2, 35392 Giessen, Germany
E-mail: malcher@informatik.uni-giessen.de

Carlo Mereghetti · Beatrice Palano
DSI, Università degli Studi di Milano, via Comelico 39/41, 20135 Milano, Italy
E-mail: {mereghetti, palano}@dsi.unimi.it

Important first-order logics for language description are FO$[+1]$, FO$[<]$ and FO$[+]$. All these logics are used to express properties of words, and their variables range over word positions. Along with the usual predicates $Q_a(x)$ holding true whenever the letter at position x is a, and equality, they are provided with the predicates $x+1=y$, $x<y$ and $x+y=z$, respectively. A further step in incrementing the descriptive power was made two decades ago by introducing the notion of counting quantifier (see, e.g., [1,6,17]) which roughly speaking enables us to count the number of values satisfying a given formula. These new quantifiers were modeled after the majority function used in circuit complexity and may not be confused with the modular quantifiers whose descriptive power is much weaker.

It is well-known (see, e.g., [18]) that FO$[+1]$ characterizes the class of locally threshold testable languages, while FO$[<]$ characterizes the wider class of star-free languages, this latter class being itself strictly contained in that of regular languages. The class of languages described in FO$[+]$ contains all the star-free languages, not all the regular languages and also contains nonregular languages. Presently, no precise definition of the class of languages characterized by FO$[+]$ is known. Thus, it is natural to investigate the possibility of representing relevant subclasses. Our first result concerns the important subclass of the bounded languages which are definable in FO$[+]$, based on the well-known notion of semilinear languages introduced by Ginsburg and Spanier in 1964 [9].

Theorem. *A bounded language is definable in* FO$[+]$ *if and only if it is semilinear.*

This is particularly interesting since such a logical characterization of bounded semilinear languages complements the known characterizations by formal grammars (e.g., simple matrix grammars [11]) and automata (e.g., certain variants of multi-head finite automata and multi-head pushdown automata [12]).

The second part of this paper is concerned with another type of issue which, we think, was only marginally considered by people working in parallel complexity and which is more relevant in the theory of languages. Indeed, we investigate the closure properties of the classes of languages defined in various logics, see Figure 1, where FOC$[+]$ refers to FO$[+]$ augmented with counting quantifiers. Some entries of the table were already known, at least implicitly. Two remain unanswered.

Table 1 Closure properties.

	FO$[+1]$	FO$[<]$	FO$[+]$	FOC$[+]$
length-preserving morphism	no	no	no	no
length-preserving inverse morphism	yes	yes	yes	yes
inverse morphism	no	yes	no	yes
concatenation	no	yes	yes	yes
shuffle	no	no	no	?
disjoint shuffle	no	yes	no	yes
Kleene star	no	no	no	yes if $\mathrm{TC}^0=\mathrm{NC}^1$
quotient	yes	yes	yes	yes
conjugate image	yes	yes	yes	yes
commutative image	no	no	no	?
reversal	yes	yes	yes	yes

Our characterization and closure properties lead us to results on the logical definability of the meaningful class of the Dyck languages [5,10]. It is known from [1] that the Dyck languages can be described by FOC[+]. Moreover, from [14], one may easily get that the Dyck languages cannot be described in FO[+]. Here, we give a new proof of this latter result relying on logics.

2 Preliminaries

The set of natural numbers is here denoted by $\mathbb{N}$. We assume basic notions on formal language theory [10]. Given an alphabet Σ, we denote by Σ^* the set of words on Σ, including the empty word ε. We denote by $|x|$ the length of a word $x \in \Sigma^*$ and by Σ^i the set of words of length i, with $\Sigma^0 = \{\varepsilon\}$. We let $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$. For any $x, w \in \Sigma^*$, we let $|x|_w$ be the number of occurrences of the word w in x. A *language* on Σ is any subset of Σ^*.

We assume familiarity with traditional and threshold circuits as computational models to study the parallel complexity of problems (see, e.g., [15,19]). We recall that NC^k (AC^k) is the class of problems solved by families of bounded (unbounded) fan-in AND/OR/NOT-circuits of polynomial size and $O(\log^k n)$ depth. The class TC^0 contains those problems solved by constant depth threshold circuits of polynomial size. We have the inclusions AC^0 $\subset$ TC^0 $\subseteq$ NC^1, and the latter inclusion is widely considered to be proper. It is known that regular languages lie in NC^1 [16], but not in TC^0 unless TC^0 = NC^1 [3]. In [13], further interesting subclasses of context-free languages are shown to be in NC^1, such as the class of the Dyck languages over an arbitrary number of parentheses, and the class of bounded semilinear languages (see Sections 3 and 5 for a definition of these languages).

The connection between circuit complexity issues and first-order logic formalisms for language description is extensively covered in [18]. In these formalisms, the words over Σ are represented as first-order structures in the signature $\langle \{Q_a\}_{a\in\Sigma}, \{P_i\}_{1\le i\le m}, \texttt{last}\rangle$, so that the structure for a word w of length n has universe $\{1,\ldots,n\}$, Q_a is the unary predicate holding true for $1 \le j \le n$ if and only if the jth letter of w is a, the P_i's are numerical predicates of different arities (e.g., $x < y$ or $x = y+z$) and $\texttt{last}$ is the constant n. In fact, all logics considered in the sequel assume the predicates Q_a, the numerical predicate $x = y$, and the constant $\texttt{last}$. Yet, they differ on the set $\mathscr{Z}$ of the assumed additional numerical predicates, e.g.: $+1$ for the immediate successor, $<$ for the usual ordering on the nonnegative integers, and $+$ for the ternary predicate $x = y+z$. In this way, we use the notation FO[$\mathscr{Z}$], where FO stands for first-order quantification. The formulas from FO[$\mathscr{Z}$] are defined in the usual way, i.e.: every atomic predicate is a formula; if φ_1 and φ_2 are formulas, then $\varphi_1 \wedge \varphi_2$, $\varphi_1 \vee \varphi_2$ and $\neg\varphi_1$ are formulas; if $\varphi(x_1,\ldots,x_n)$ is a formula whose free variables are $x_1,\ldots,x_n$, then $\exists x_i\, \varphi(x_1,\ldots,x_n)$ and $\forall x_i\, \varphi(x_1,\ldots,x_n)$, with $1 \le i \le n$, are formulas.

The notion of *counting quantifiers* is less known and deserves a precise definition. We restrict ourselves to the unary version. Given a formula $\phi(x_1,\ldots,x_n)$ and a variable y, $\exists^y_{x_i}\phi(x_1,\ldots,x_n)$ is a formula with free variables $y, x_1,\ldots,x_{i-1}$, $x_{i+1},\ldots,x_n$. It is true for the assignment b to y and a_j to x_j, $j \ne i$, if there exist $b > 0$ values a for x_i such that $\phi(a_1,\ldots,a_{i-1},a,a_{i+1},\ldots,a_n)$ holds true. In other words it is equivalent to the formula

$$(1 \le y \le \texttt{last}) \wedge (y = |\{x \mid \phi(x_1,\ldots,x_{i-1},x,x_{i+1},\ldots,x_n)\}|).$$

In particular, it there exists no such a, then the formula is a contradiction.

Formulas are meant to specify languages. Indeed, if φ is a sentence (a formula without free variables), we let L_φ be the set of all words satisfying φ. Formally, we write $L_\varphi = \{w \in \Sigma^* \mid w \models \varphi\}$. In this case, we say that L_φ is the language *defined* (or *described*) by φ. We denote by $\mathscr{L}$(FO[$\mathscr{Z}$]) the class of languages *definable* in FO[$\mathscr{Z}$], i.e., $\mathscr{L}$(FO[$\mathscr{Z}$]) =

$\{L \subseteq \Sigma^* \mid L = L_\varphi$, for some sentence $\varphi \in \mathrm{FO}[\mathscr{Z}]\}$. Several classes of languages have been logically characterized. For instance, FO$[+1]$ (resp., FO$[<]$) is the first-order logic with numerical predicate $+1$ (resp., $<$). It is well-known that $\mathscr{L}(\mathrm{FO}[+1])$ is the class of locally threshold testable languages, while $\mathscr{L}(\mathrm{FO}[<])$ is the class of star-free languages. No formal language characterization for $\mathscr{L}(\mathrm{FO}[+])$ is currently known. We denote by FOC the set of first-order quantifiers together with the counting quantifier '$\exists^y$'. In the sequel, we use the predicate symbol '$*$' for the natural multiplication. The following strict inclusions are well-known (see [17], for more details):

$$\mathscr{L}(\mathrm{FO}[+1]) \subset \mathscr{L}(\mathrm{FO}[<]) \subset \mathscr{L}(\mathrm{FO}[+]) \subset \begin{array}{c} \mathscr{L}(\mathrm{FO}[+,*]) = \mathrm{AC}^0 \\ \mathscr{L}(\mathrm{FOC}[<]) = \mathscr{L}(\mathrm{FOC}[+]) \end{array} \subset \mathrm{TC}^0 \subseteq \mathrm{NC}^1.$$

3 Bounded Languages

In this section, we exhibit a relevant class of languages contained in $\mathscr{L}(\mathrm{FO}[+])$, and show that, *restricted to bounded languages,* FO$[+]$ *characterizes the semilinear languages.*

We recall that a set $X \subseteq \mathbb{N}^m$ is *linear* whenever, for some integer $r \in \mathbb{N}$, there exist vectors $v_0, \ldots, v_r \in \mathbb{N}^m$ such that $X = v_0 + \sum_{t=1}^{r} \mathbb{N} v_t$. A *semilinear set* is a finite union of linear sets. Let $w_1, \ldots, w_m$ be a sequence of words in Σ^* where the same word may have several occurrences. We consider the *natural embedding of* $\mathbb{N}^m$ *into* Σ^* (we shall simply say "the natural embedding") *relative to this sequence* $w_1, \ldots, w_m$ as the mapping $\chi : \mathbb{N}^m \to \Sigma^*$ defined by $\chi(n_1, \ldots, n_m) = w_1^{n_1} \cdots w_m^{n_m}$. A language $L \subseteq \Sigma^*$ is *bounded* whenever $L = \chi(X)$ holds for some $X \subseteq \mathbb{N}^m$. It is *letter bounded* whenever all w_i's are letters. Moreover, L is bounded *(linear) semilinear* whenever X is a (linear) semilinear set. Rigorously speaking, we should always specify the sequence $w_1, \ldots, w_m$ with the mapping χ. However, the context should clearly determine which sequence is assumed.

We first prove the result under the hypothesis that the language is letter bounded, and then we extend it to arbitrary bounded languages.

3.1 A preliminary reduction

Notice that, without loss of generality, we may assume $w_i \neq w_{i+1}$ for every $1 \leq i < m$. Indeed, it clearly suffices to consider the case where X is linear, i.e., of the form $v_0 + \sum_{t=1}^{r} \mathbb{N} v_t$. Suppose there exists $1 \leq i < m$ such that $w_i = w_{i+1}$, and denote with $v_{t,j}$ the jth component of the vector $v_t \in \mathbb{N}^m$. With each v_t, for $1 \leq t \leq r$, we associate the vector $v'_t \in \mathbb{N}^{m-1}$ defined by

$$v'_{t,j} = \begin{cases} v_{t,j} & \text{if } j < i \\ v_{t,j} + v_{t,j+1} & \text{if } j = i \\ v_{t,j+1} & \text{otherwise.} \end{cases}$$

By letting χ' be the natural embedding of $\mathbb{N}^{m-1}$ into $w_1^* \cdots w_i^* w_{i+2}^* \cdots w_m^*$, we get $L = \chi'(X')$ with the linear set $X' = v'_0 + \sum_{t=1}^{r} \mathbb{N} v'_t$.

Some closure properties of the bounded semilinear languages are obtained as technical adaptations of the closure properties of the semilinear sets in $\mathbb{N}^m$, as we shall discuss in Proposition 1. The following is a technical property which deals with the problem of the occurrence of the empty word among $w_1, \ldots, w_m$.

Lemma 1 *Every bounded semilinear subset of Σ^* is a finite union of linear subsets in $w_1^+ \cdots w_m^+$, with $w_i \neq \varepsilon$ for $i = 1, \ldots, m$, and possibly of the subset containing only the empty word.*

Proof Indeed, the sets of the form $M_1 \times \cdots \times M_m$, where $M_i = \{0\}$ or $M_i = \mathbb{N} \setminus \{0\}$, are linear and thus so are all intersections $X \cap M_1 \times \cdots \times M_m$, with X a linear set. Let $\emptyset \subset I \subseteq \{1, \ldots, m\}$ be the set of indices i such that $M_i = \mathbb{N} \setminus \{0\}$, let π_I be the projection of $\mathbb{N}^m$ onto $\prod_{i \in I} M_i$, and let χ' be the unique natural embedding satisfying $\chi(v) = \chi'(\pi_I(v))$, for all $v \in \mathbb{N}^m$, as defined at the beginning of this section. Then, we get

$$\chi(X \cap M_1 \times \cdots \times M_m) = \chi'\left(\pi_I(X) \cap \prod_{i \in I} M_i\right).$$

Whenever $I = \emptyset$, we have $\chi(X \cap M_1 \times \cdots \times M_m) = \{\varepsilon\}$. □

3.2 The Letter Case

A morphism of a free monoid into another is *nonincreasing* if the image of a letter is a letter or the empty word. A *length-preserving substitution* is defined by a mapping $h : \Sigma \to 2^\Sigma \setminus \{\emptyset\}$, and assigns to the word $a_1 \cdots a_n$ the set of words $h(a_1) \cdot \ldots \cdot h(a_n)$. It extends to subsets of words in the usual way.

Proposition 1 *If $L, L' \subseteq a_1^* \cdots a_m^*$ are letter bounded semilinear languages and if f is a nonincreasing morphism, then $L \cup L'$, $L \setminus L'$, LL', and $f(L)$ are letter bounded semilinear. Furthermore, if h is a length-preserving substitution, then $L \cap h(L')$ is letter bounded semilinear.*

Proof The first three properties are consequences of the results in [8]. Concerning nonincreasing morphisms, observe that they are a composition of morphisms of two types: those renaming a letter and leaving all other letters invariant and those sending a letter to the empty word and leaving all other letters invariant. In the former case, the result follows directly from the above observation concerning the non-repetition of two consecutive words. In the latter case, assume the morphism f satisfies $f(a) = \varepsilon$ and $f(c) = c$ for all $c \in \Sigma \setminus \{a\}$. Let I the subset of indices $i \in \{1, \ldots, m\}$ such that $a_i = a$, and denote by π the morphism of $\mathbb{N}^m$ into $\mathbb{N}^{m-|I|}$ assigning to any m-tuple the $(m - |I|)$-tuple obtained by ignoring the components whose positions are in I. If $L = \chi(X)$ then $f(L) = \chi'(\pi(X))$, where χ' is the obvious restriction of χ, and we apply the closure property under morphism of the semilinear sets [8].

The last statement follows from the fact that the intersection of a bounded semilinear language with the image of a bounded semilinear language under a substitution by regular sets is bounded semilinear [11, Thm 5.5]. Observe that $h(L')$ is not bounded in general, which is why we consider the intersection with a bounded language. □

We are now going to show that every letter bounded language is in $\mathscr{L}(\text{FO}[+])$ if and only if it is semilinear. We start with the “if” part.

Theorem 1 *The class of letter bounded semilinear languages is in* $\mathscr{L}(\text{FO}[+])$.

Proof It clearly suffices to prove the result for letter bounded linear languages, i.e., languages of the form $L = \chi(X)$ where $X \subseteq \mathbb{N}^m$ is linear and χ is the natural embedding of $\mathbb{N}^m$ into $a_1^* \cdots a_m^*$, for some sequence $a_1, \ldots, a_m$ of letters from Σ.

By Lemma 1, we may assume that $L \subseteq a_1^+ \cdots a_m^+$. We give an FO[+] formula defining the letter bounded linear language $L = \chi(X)$, with $X = v_0 + \sum_{t=1}^{r} \mathbb{N} v_t$. By definition, a word $w \in \Sigma^*$ belongs to L if and only if:

(i) w is in $a_1^+ \cdots a_m^+$, and
(ii) for some $\alpha_1, \dots, \alpha_r \in \mathbb{N}$, it holds $|w|_{a_j} = v_{0,j} + \sum_{t=1}^{r} \alpha_t v_{t,j}$ for $1 \leq j \leq m$, where $v_{t,j}$ denotes the jth component of the vector v_t.

Let the variables $y_1, \dots, y_m$ be interpreted as the number of occurrences of the letters $a_1, \dots, a_m$ in w. Then, an FO[+] formula defining L is of the form

$$\exists y_1 \cdots \exists y_m \; \psi_1(y_1, \dots, y_m) \wedge \psi_2(y_1, \dots, y_m),$$

where ψ_1 expresses condition (i) and ψ_2 expresses condition (ii). Using natural abbreviations in order to keep the formula readable, we have

$$\begin{aligned} \psi_1(y_1, \dots, y_m) \equiv{} & (y_1 + \cdots + y_m = \mathtt{last}) \wedge \\ & \forall z \Big((z \leq y_1 \Rightarrow Q_{a_1}(z)) \wedge \bigwedge_{i=2}^{m} \Big(\sum_{h=1}^{i-1} y_h < z \leq \sum_{h=1}^{i} y_h \Rightarrow Q_{a_i}(z) \Big) \Big). \end{aligned}$$

Concerning condition (ii), by denoting with z_j the variables interpreted as the coefficients α_j, we have

$$\psi_2(y_1, \dots, y_m) \equiv \exists z_1 \cdots \exists z_r \Big(\bigwedge_{1 \leq j \leq m} \Big(y_j = v_{0,j} + \sum_{t=1}^{r} \sum_{s=1}^{v_{t,j}} z_t \Big) \Big).$$

□

The converse of Theorem 1 goes by structural induction on FO[+] formulas. Indeed, we consider not only sentences (i.e., formulas with only bound variables) but more generally formulas with free variables, and define what it means for such formulas to be satisfied by some model. We utilize the usual trick which consists of augmenting the letters of the alphabet Σ with a new component specifying subsets of free variables: by so doing, we encode the value of the free variables in the model. More precisely, a formula ϕ over a set $\mathscr{V}$ of free variables is interpreted on *$\mathscr{V}$-structures*, i.e., words of the form $u = (\sigma_1, \mathscr{V}_1) \cdots (\sigma_n, \mathscr{V}_n)$ over the alphabet $\Sigma \times 2^{\mathscr{V}}$, with: (i) $\mathscr{V}_i \subseteq \mathscr{V}$, (ii) $\mathscr{V}_i \cap \mathscr{V}_j = \emptyset$ for $i \neq j$, (iii) $\bigcup_{i=1}^{n} \mathscr{V}_i = \mathscr{V}$. We let $\mathscr{S}_{|\mathscr{V}|} \subseteq (\Sigma \times 2^{\mathscr{V}})^*$ denote the set of all $\mathscr{V}$-structures.

Let us now explain what it means for a $\mathscr{V}$-structure to satisfy a formula with free variables (the figure below should facilitate the intuition). To fix ideas, let $\phi(x_1, \dots, x_k)$ be a formula and $\mathscr{V} = \{x_1, \dots, x_k\}$ be the set of its free variables. We say that the $\mathscr{V}$-structure $u = (\sigma_1, \mathscr{V}_1) \cdots (\sigma_n, \mathscr{V}_n) \in \mathscr{S}_{|\mathscr{V}|}$ satisfies $\phi(x_1, \dots, x_k)$, and we write $u \models \phi(x_1, \dots, x_k)$, if $\phi(p_1, \dots, p_k)$ holds true in the model $\sigma_1 \cdots \sigma_n$ where, for $1 \leq i \leq k$, the integer $1 \leq p_i \leq n$ is the unique position of the $\mathscr{V}$-structure u such that $x_i \in \mathscr{V}_{p_i}$. For instance, the following $\mathscr{V}$-structure

a	b	a	a	b	b
$\emptyset$	$\{x_2\}$	$\{x_3, x_4\}$	$\emptyset$	$\{x_1\}$	$\emptyset$

satisfies the formula

$$Q_b(x_1) \wedge Q_b(x_2) \wedge \neg Q_b(x_3) \wedge (x_2 < x_3) \wedge (x_3 < x_1) \wedge (x_2 < x_4) \wedge (x_4 < x_1).$$

The language defined by ϕ is the set $L_{\phi,\mathscr{V}} = \{u \in \mathscr{S}_{|\mathscr{V}|} \mid u \models \phi\}$. If ϕ is a sentence, i.e., a formula without free variables, then $L_\phi = L_{\phi,\emptyset} = \{w \in \Sigma^* \mid w \models \phi\}$, as recalled in Section 2.

The following lemma, which is useful in the proof of the main result, shows the letter boundedness and semilinearity of a language of $\mathscr{V}$-structures. Given a $\mathscr{V}$-structure $u = (\sigma_1, \mathscr{V}_1) \cdots (\sigma_n, \mathscr{V}_n)$, we define $\pi(u) = \sigma_1 \cdots \sigma_n$. Then:

Lemma 2 *Let $a_1, \ldots, a_m$ be a sequence of letters from an alphabet Σ and let $\mathscr{V}$ be a set of free variables with $|\mathscr{V}| = k$. The language*

$$B_k = \{u \in \mathscr{S}_k \mid \pi(u) \in a_1^* \cdots a_m^*\}$$

is letter bounded semilinear.

Proof The set $a_1^* \cdots a_m^*$ is the (finite) union of subsets of the form $a_{i_1}^+ \cdots a_{i_s}^+$ where $0 \leq s \leq m$ and $i_1, \ldots, i_s$ is a subsequence of $1, \ldots, m$. So, we are reduced to prove that the set $B'_k = \{u \in \mathscr{S}_k \mid \pi(u) \in a_1^+ \cdots a_m^+\}$ is letter bounded semilinear.

We claim that B'_k is a finite union of letter bounded linear languages. A subset in this union is specified by the choice of a sequence of nonempty subsets $\mathscr{V}_1, \ldots, \mathscr{V}_\ell$ defining a decomposition of $\mathscr{V}$ and a choice of letters from $a_1, \ldots, a_m$ which are associated with the $\mathscr{V}_i$'s. For instance, let $k = 3$ and $m = 3$, i.e., we are considering the words in $a_1^+ a_2^+ a_3^+$ and the formula has 3 free variables x_1, x_3, x_3. Consider the decomposition $\mathscr{V} = \mathscr{V}_1 \cup \mathscr{V}_2$ with $\mathscr{V}_1 = \{x_1, x_3\}$ and $\mathscr{V}_2 = \{x_2\}$. Associate $\mathscr{V}_1$ with a_1 and $\mathscr{V}_2$ with a_3. Then, the associated subset is

$$(a_1, \emptyset)^* (a_1, \mathscr{V}_1)(a_1, \emptyset)^* (a_2, \emptyset)^* (a_3, \emptyset)^* (a_3, \mathscr{V}_2)(a_3, \emptyset)^*,$$

which indicates that the positions of the interpretations of the variables x_1 and x_3 are inside the factor of the word labeled by a_1, while the interpretation of the variable x_2 is inside the factor of the word labeled by a_3.

Formally, consider a sequence of the form $(i_1, \mathscr{V}_1), \ldots, (i_\ell, \mathscr{V}_\ell)$ satisfying the following conditions:

- $1 \leq i_1 < \cdots < i_\ell \leq m$,
- $\mathscr{V}_{i_\alpha} \neq \emptyset$, for $1 \leq \alpha \leq \ell$,
- $\bigcup_{\alpha=1}^{\ell} \mathscr{V}_{i_\alpha} = \mathscr{V}$ and $\mathscr{V}_{i_\alpha} \cap \mathscr{V}_{i_\beta} = \emptyset$, for $1 \leq \alpha < \beta \leq \ell$.

Set $i_0 = 1$ and $i_{\ell+1} = m$, and define

$$L_\alpha = (a_{i_{\alpha-1}}, \emptyset)^* \cdots (a_{i_\alpha}, \emptyset)^* (a_{i_\alpha}, \mathscr{V}_\alpha)(a_{i_\alpha}, \emptyset)^* \cdots (a_{i_{\alpha+1}}, \emptyset)^*.$$

Then, each of these L_α is letter bounded linear and thus, by Proposition 1, the product $L_1 \cdots L_\ell$ is letter bounded linear as well. Since B'_k is a finite union of such languages, it is letter bounded semilinear. □

Hereafter, for the sake of conciseness, given an alphabet Σ we let $Q_\Sigma(x)$ stand for $\bigvee_{\sigma \in \Sigma} Q_\sigma(x)$. We obtain the converse of Theorem 1 as a corollary of

Theorem 2 *For every sequence $a_1, \ldots, a_m$ of letters in Σ and every language $L \subseteq \Sigma^*$ in $\mathscr{L}(\mathrm{FO}[+])$, the language $L \cap a_1^* \cdots a_m^*$ is semilinear.*

Proof Let ϕ be an FO[+] formula with $\mathscr{V} = \{x_1, \ldots, x_k\}$ free variables. To prove the result, it is enough to show that $L_{\phi, \mathscr{V}} \cap B_k$ is semilinear, where $B_k = \{u \in \mathscr{S}_k \mid \pi(u) \in a_1^* \cdots a_m^*\}$ is the letter bounded semilinear language in Lemma 2. We shall use the structural induction on ϕ, starting from atomic predicates and then consider more complex formulas.

- $\phi \equiv Q_a(x)$: If a does not occur in the sequence $a_1, \dots, a_m$, then $L_{\phi,\mathscr{V}} \cap B_1 = \emptyset$, so we assume $a = a_i$ for some $1 \leq i \leq m$. We have
$$L_{\phi,\{x\}} \cap B_1 = (a_1,\emptyset)^* \cdots (a_i,\emptyset)^*(a_i,\{x\})(a_i,\emptyset)^* \cdots (a_m,\emptyset)^*,$$
which is clearly letter bounded linear.
- $\phi \equiv (x_1 + x_2 = x_3)$: We have to show that $L_{\phi,\{x_1,x_2,x_3\}} \cap B_3$ is semilinear. The formula ϕ is equivalent to $(\phi \wedge (x_1 < x_2)) \vee (\phi \wedge (x_1 = x_2)) \vee (\phi \wedge (x_1 > x_2))$. We prove that the language $L_{\phi_<,\{x_1,x_2,x_3\}} \cap B_3$ with $\phi_< \equiv \phi \wedge (x_1 < x_2)$ is letter bounded linear. The two other cases are treated similarly. For $W \subseteq \{x_1,x_2,x_3\}$, let $A_W = \{(\sigma,W) \mid \sigma \in \{a_1,a_2,\dots,a_m\}\}$. Then
$$L_{\phi_<,\{x_1,x_2,x_3\}} = \{A_\emptyset{}^\alpha A_{\{x_1\}} A_\emptyset{}^\beta A_{\{x_2\}} A_\emptyset{}^\alpha A_{\{x_3\}} A_\emptyset{}^\gamma \mid \alpha,\beta,\gamma \in \mathbb{N}\}.$$
This language is the image under the length-preserving substitution defined by $h(a) = A_\emptyset$, $h(b) = A_{\{x_1\}}$, $h(c) = A_{\{x_2\}}$ and $h(d) = A_{\{x_3\}}$ of the letter bounded linear language $L' = \{a^\alpha b a^\beta c a^\alpha d^\gamma \mid \alpha,\beta,\gamma \in \mathbb{N}\}$. So
$$L_{\phi_<,\{x_1,x_2,x_3\}} \cap B_3 = h(L') \cap B_3,$$
and the result follows from Lemma 2 and Proposition 1.
- $\phi \equiv \neg\psi$: We have
$$\begin{aligned} L_{\phi,\mathscr{V}} \cap B_k &= L_{\neg\psi,\mathscr{V}} \cap B_k = (\mathscr{S}_k \setminus L_{\psi,\mathscr{V}}) \cap B_k = (\mathscr{S}_k \cap L^c_{\psi,\mathscr{V}}) \cap B_k \\ &= L^c_{\psi,\mathscr{V}} \cap B_k = B_k \cap (L^c_{\psi,\mathscr{V}} \cup B^c_k) = B_k \setminus (L_{\psi,\mathscr{V}} \cap B_k). \end{aligned}$$
By inductive hypothesis, we have that $L_{\psi,\mathscr{V}} \cap B_k$ is letter bounded semilinear. The result follows from Lemma 2 and Proposition 1.
- $\phi \equiv \psi_1 \wedge \psi_2$: We first transform ψ_1 and ψ_2 into equivalent formulas, each over the same set of free variables, say $\mathscr{V}$. To this purpose, let $\mathscr{W}_1$ and $\mathscr{W}_2$ be the set of free variables of ψ_1 and ψ_2, respectively, so that $\mathscr{V} = \mathscr{W}_1 \cup \mathscr{W}_2$. Define $\hat{\psi}_1 \equiv \psi_1 \wedge \bigwedge_{x \in \mathscr{V} \setminus \mathscr{W}_1} Q_\Sigma(x)$ and $\hat{\psi}_2 \equiv \psi_2 \wedge \bigwedge_{x \in \mathscr{V} \setminus \mathscr{W}_2} Q_\Sigma(x)$. Clearly, ϕ is equivalent to $\hat{\psi}_1 \wedge \hat{\psi}_2$. We have
$$\begin{aligned} L_{\phi,\mathscr{V}} \cap B_k &= L_{\hat{\psi}_1 \wedge \hat{\psi}_2,\mathscr{V}} \cap B_k = L_{\hat{\psi}_1,\mathscr{V}} \cap L_{\hat{\psi}_2,\mathscr{V}} \cap B_k \\ &= (L_{\hat{\psi}_1,\mathscr{V}} \cap B_k) \cap (L_{\hat{\psi}_2,\mathscr{V}} \cap B_k). \end{aligned}$$
By inductive hypothesis, $L_{\psi_1,\mathscr{W}_1} \cap B_{|\mathscr{W}_1|}$ and $L_{\psi_2,\mathscr{W}_2} \cap B_{|\mathscr{W}_2|}$ are letter bounded semilinear. Let h be the length-preserving substitution which assigns to every letter $(a,W) \in \mathscr{S}_{|\mathscr{W}_1|}$ the set of letters $(a,V) \in \mathscr{S}_k$ where $V = W \cup A$, for $A \subseteq \mathscr{W}_2 \setminus \mathscr{W}_1$. Then, we have
$$L_{\hat{\psi}_1,\mathscr{V}} \cap B_k = h(L_{\psi_1,\mathscr{W}_1} \cap B_{|\mathscr{W}_1|}) \cap B_k.$$
The result follows from Lemma 2 and Proposition 1.
- $\phi \equiv \exists x_{k+1}\ \psi(x_1,\dots,x_k,x_{k+1})$: Let the nonincreasing morphism $\Pi : \mathscr{S}_{k+1} \to \mathscr{S}_k$ be defined as $\Pi(a,V) = (a,V \setminus \{x_{k+1}\})$. E.g., if $u = (a,\emptyset)(a,\{x_3\})(b,\{x_1\})(b,\{x_2\}) \in \mathscr{S}_3$, we have $\Pi(u) = (a,\emptyset)(a,\emptyset)(b,\{x_1\})(b,\{x_2\}) \in \mathscr{S}_2$. We extend Π to subsets of $\mathscr{S}_{k+1}$ in the usual way. Notice that $B_k = \Pi(B_{k+1})$. We have
$$\begin{aligned} L_{\phi,\mathscr{V}} \cap B_k &= L_{\exists x_{k+1}\ \psi(x_1,\dots,x_{k+1}),\mathscr{V}} \cap B_k = \Pi(L_{\psi(x_1,\dots,x_{k+1}),\mathscr{V} \cup \{x_{k+1}\}}) \cap \Pi(B_{k+1}) \\ &= \Pi(L_{\psi(x_1,\dots,x_{k+1}),\mathscr{V} \cup \{x_{k+1}\}} \cap B_{k+1}). \end{aligned}$$
The last equality follows from the saturation of B_{k+1} relative to Π, i.e., from the fact that $\Pi(u) = \Pi(v)$ and $u \in B_{k+1}$ implies $v \in B_{k+1}$. By inductive hypothesis, we have that $L_{\psi(x_1,\dots,x_{k+1}),\mathscr{V} \cup \{x_{k+1}\}} \cap B_{k+1}$ is letter bounded semilinear, and the result follows from Proposition 1. □

In conclusion, from Theorems 1 and 2, we get

Theorem 3 *A letter bounded language is semilinear if and only if it belongs to* $\mathscr{L}(\mathrm{FO}[+])$.

3.3 From Letters to Words

We show how to extend the result from letter bounded to word bounded languages. From now on, the term bounded means word bounded unless otherwise stated.

Theorem 4 *Let L be a bounded language. Then, L is semilinear if and only if L belongs to* $\mathscr{L}(\mathrm{FO}[+])$.

Proof By Lemma 1, we assume without loss of generality that the semilinear bounded language we start with satisfies the further condition

$$L = \{w_1^{r_1} \cdots w_m^{r_m} \mid (r_1, \dots, r_m) \in X\},$$

where X consists of m-tuples with nonzero components. We also introduce the alphabet $\{a_1, \dots, a_m\}$ of m new distinct symbols, where a_i is in correspondence with w_i. Observe that there might be several occurrences of the same word. We construct an FO[+] formula which defines L. Denoting by Y the componentwise product $(|w_1|, \dots, |w_m|)X$, the subset

$$L' = \{a_1^{|w_1|r_1} \cdots a_m^{|w_m|r_m} \mid (|w_1|r_1, \dots, |w_m|r_m) \in Y\}$$

is linear and therefore there exists an FO[+] formula ϕ' for L'. Let us introduce the following predicates:

- $\psi_{a_i}(x, y, x') \equiv x \le y \le x'$, $i = 1, \dots, m$.
- $\gamma_i(x, x')$ which holds true whenever the factor of a word between the position x and x', both ends included, belongs to w_i^+. Formally, set $w_i = w_{i,1} \dots w_{i,k_i}$. We define

$$\gamma_i(x,x') \equiv \bigwedge_{1 \le j \le k_i} Q_{w_{i,j}}(x - 1 + j) \wedge \\ \forall y \Big((x \le y \le x' - k_i) \wedge \Big(\bigwedge_{1 \le j \le k_i} Q_{w_{i,j}}(y) \Rightarrow \bigwedge_{1 \le j \le k_i} Q_{w_{i,j}}(y + k_i)\Big)\Big).$$

Then, the FO[+] formula for L is as follows:

$$\exists x_1 \cdots \exists x_{m+1} \Big((1 = x_1 < \cdots < x_{m+1} = \mathtt{last}) \wedge \gamma_1(x_1, x_2) \wedge \bigwedge_{1 < i \le m} \gamma_i(x_i + 1, x_{i+1}) \wedge \phi\Big),$$

where ϕ is obtained from ϕ' by substituting $\psi_{a_1}(x_1, y, x_2)$ for $Q_{a_1}(y)$ and, for $1 < i \le m$, $\psi_{a_i}(x_i + 1, y, x_{i+1})$ for $Q_{a_i}(y)$.

Conversely, consider a bounded language L which is defined by the formula ϕ in FO[+]. The idea is to transform every word $w_1^{r_1} \cdots w_m^{r_m}$ into $a_1^{|w_1|r_1} \cdots a_m^{|w_m|r_m}$ and to show that the subset L' of $a_1^* \cdots a_m^*$ thus obtained is definable in FO[+]. Without loss of generality, here we consider the case $r_1, \dots, r_m > 0$:

$$\exists x_1 \cdots \exists x_{m+1} \Big((1 = x_1 < \cdots < x_{m+1} = \mathtt{last}) \wedge \forall y \, (x_1 \le y \le x_2 \Rightarrow Q_{a_1}(y)) \wedge \\ \Big(\bigwedge_{1 < i \le m} \forall y \, (x_i < y \le x_{i+1} \Rightarrow Q_{a_i}(y))\Big) \wedge \phi'\Big),$$

where the formula ϕ' is obtained from the formula ϕ by substituting every occurrence of $Q_\sigma(y)$ by the following predicate. Let $I \subseteq \{1,\ldots,m\}$ be such that $i \in I$ implies that w_i contains the symbol σ, and let $|w_i| = k_i$. Then:

$$\psi_\sigma(y) \equiv \bigvee_{i\in I} \bigvee_{w_{i,j}=\sigma} \exists z\,(y = x_i + 1 + j + k_i z).$$

By Theorem 3, the subset L' is semilinear, therefore so is the subset L as can be readily verified. This completes the proof. □

4 Closure Properties

In this section, we investigate the closure properties of the classes $\mathscr{L}(\text{FO}[+1])$, $\mathscr{L}(\text{FO}[<])$, $\mathscr{L}(\text{FO}[+])$ and $\mathscr{L}(\text{FOC}[+])$ under various operations. The results of the previous section help us to show some negative closure results. On the other hand, also some positive closure results are obtained. The complete picture is displayed in Figure 1 of the Introduction.

The following notation will be used in this section: Let $\phi(x_1,\ldots,x_r)$ be a formula with $x_1,\ldots,x_r$ free variables. We write $(w,a_1,\ldots,a_r) \models \phi$ whenever $w \models \phi(a_1,\ldots,a_r)$.

4.1 Morphisms

We need a more precise classification of the morphisms. We recall that a morphism $h : \Sigma^* \longrightarrow \Delta^*$ of a free monoid into another is entirely defined by the image of each letter of Σ. It is an elementary result that h can be expressed as a composition of morphisms g of one of the following types:

- g defines a permutation on the symbols of Σ and is called *permutation* morphism.
- g identifies two letters and leaves all other letters invariant, i.e., there exist $a,b \in \Sigma$ such that $g(a) = g(b) = a$ and $g(c) = c$ for all $c \neq a,b$. Call this morphism *identifying*.
- g deletes some letter and leaves all other letters unchanged, i.e, there exists $a \in \Sigma$ such that $g(a) = \varepsilon$ and $g(b) = b$ for all $b \neq a$. This morphism is called *erasing*.
- there exists a letter $a \in \Sigma$ such that $g(a) = ac$ and $g(b) = b$ for all $b \neq a$. Let us call the morphism *growing*.

Consistently with the definition given at the beginning of Section 3.2, we say that a morphism $h : \Sigma^* \to \Delta^*$ is *length-preserving* whenever $|h(w)| = |w|$, for every $w \in \Sigma^*$. We start by showing that none of the classes considered are closed under length-preserving homomorphisms.

Proposition 2 *The classes* $\mathscr{L}(\text{FO}[+1])$*,* $\mathscr{L}(\text{FO}[<])$*,* $\mathscr{L}(\text{FO}[+])$ *and* $\mathscr{L}(\text{FOC}[+])$ *are not closed under length-preserving homomorphism.*

Proof Consider $\Sigma = \{a,b\}$ and the language $L = (ab)^+$. Notice that $(ab)^+ = a\Sigma^* \cap \Sigma^* b \setminus \Sigma^*\{a^2,b^2\}\Sigma^*$ holds, which implies L is a local set and belongs to $\mathscr{L}(\text{FO}[+1])$ and thus to $\mathscr{L}(\text{FO}[<])$. Now, consider the length-preserving homomorphism $h : \{a,b\}^+ \to \{a\}^+$ defined as $h(a) = h(b) = a$. We obtain $h(L) = (a^2)^+$, which does not belong to $\mathscr{L}(\text{FO}[<])$.

Concerning $\mathscr{L}(\text{FO}[+])$, we start by showing that the language

$$L = \{abab^2ab^3a\cdots ab^ia\cdots ab^ka \mid k \in \mathbb{N} \setminus \{0\}\}$$

belongs to $\mathscr{L}(\mathrm{FO}[+])$ but $h(L)$ does not belong to $\mathscr{L}(\mathrm{FOC}[+])$, where h is the length-preserving homomorphism above defined. Observe that $w \in \{a,b\}^*$ belongs to L if and only if both the following conditions hold true:

- $w = aba$ or $w \in aba\Sigma^*a$,
- if $w \neq aba$ and $w = uab^iav$, with $u,v \in \{a,b\}^*$ and $i > 0$, then there exists a suffix v' of v such that $v = b^{i+1}av'$.

These two conditions can be expressed by the following FO[+] formula

$$\begin{aligned}&Q_a(1) \wedge Q_b(2) \wedge Q_a(3) \wedge Q_a(\mathtt{last})\wedge\\ &\forall x\,[(\exists y \exists z\; Q_a(x) \wedge Q_a(x+y) \wedge Q_a(x+y+z)) \Rightarrow (\exists y\; \phi(x,y))],\end{aligned}$$

with

$$\begin{aligned}\phi(x,y) \equiv\; &Q_a(x) \wedge Q_a(x+y+1) \wedge Q_a(x+2y+3)\wedge\\ &\forall z\,(x < z < x+y+1 \Rightarrow Q_b(z))\wedge\\ &\forall z\,(x+y+1 < z < x+2y+3 \Rightarrow Q_b(z)).\end{aligned}$$

Now, we have

$$h(L) = \{a^{\frac{(p+1)(p+2)}{2}} \mid p \in \mathbb{N} \setminus \{0\}\}.$$

By Theorem 2, $h(L) \notin \mathscr{L}(\mathrm{FO}[+])$. Moreover, we know from [17] that FO[+] and FOC[+] have the same expressive power when used for unary languages. This implies that $h(L) \notin \mathscr{L}(\mathrm{FOC}[+])$ as well. □

The situation is slightly different concerning the inverse morphisms.

Proposition 3 *The class* $\mathscr{L}(\mathrm{FO}[<])$ *is closed under inverse homomorphism, but the classes* $\mathscr{L}(\mathrm{FO}[+1])$ *and* $\mathscr{L}(\mathrm{FO}[+])$ *are not. All three classes are closed under inverse length-preserving homomorphism.*

Proof The assertion for $\mathscr{L}(\mathrm{FO}[<])$ follows from the fact that it is a pseudovariety of languages [18]. Concerning the nonclosure of $\mathscr{L}(\mathrm{FO}[+1])$ and $\mathscr{L}(\mathrm{FO}[+])$, we observe that $X = \{ab\} \in \mathscr{L}(\mathrm{FO}[+1])$ and $Y = \{(a^2)^*\} \in \mathscr{L}(\mathrm{FO}[+]) \setminus \mathscr{L}(\mathrm{FO}[<])$. Consider the morphisms $h : \{a,b,c\}^* \to \{a,b\}^*$ and $g : \{a,b\}^* \to \{a\}^*$ defined by $h(a) = a$, $h(b) = b$, $h(c) = \varepsilon$ and $g(a) = a$, $g(b) = \varepsilon$. Then, $h^{-1}(X) = c^*ac^*bc^* \notin \mathscr{L}(\mathrm{FO}[+1])$ and, by [2, Cor. 4.2] which is the Crane-Beach conjecture for FO[+], $g^{-1}(Y) = \{w \in \{a,b\}^+ \mid |w|_a \bmod 2 = 0\} \notin \mathscr{L}(\mathrm{FO}[+])$.

The last statement holds in all generality. Indeed, a length-preserving homomorphism h is a composition of permutation morphisms and identifying morphisms. We thus only consider the latter case. Let $a,b \in \Sigma$, $h(a) = h(b) = a$ and $h(c) = c$ whenever $c \neq a,b$. If $X = L_\phi$ then $h^{-1}(X) = L_{\phi'}$ where ϕ' is obtained from ϕ by substituting $Q_a(x) \vee Q_b(x)$ for each occurrence of $Q_a(x)$. □

The class FOC[+] behaves very differently since it is closed under all inverse homomorphisms. We need a preliminary general result showing that under general conditions and provided the language allows counting quantifiers, it is possible to simulate a structure into another. This idea is similar to the notion of FO-reduction as exposed, e.g., in [6]. We prove that in this embedding, sums and counting quantifications in the first structure are expressible in the other. Given a total ordering, we call *rank* of an element x the number $\rho(x)$ of elements less than or equal to x, i.e., $\rho(x) = |\{y \mid y \leq x\}|$.

Proposition 4 *Let $U(x,y) \in \mathrm{FOC}[+]$ be a predicate, let $(x_1,y_1) \leq (x_2,y_2)$ be a predicate in $\mathrm{FOC}[+]$ which defines a total ordering on $\{(x,y) \mid U(x,y)\}$ and let $\rho(x,y)$ be the rank of (x,y) in this ordering. Let k be an integer such that $(a,b) \models U(x,y)$ implies $a \leq k$. Then:*

(i) There exists a formula $\theta(x_1,y_1,x_2,y_2,x_3,y_3)$ in $\mathrm{FOC}[+]$ which holds true if and only if for $(x_1,y_1),(x_2,y_2),(x_3,y_3) \models U(x,y)$ the following property holds:

$$\rho(x_1,y_1) + \rho(x_2,y_2) = \rho(x_3,y_3). \tag{1}$$

(ii) Let $\phi((x_1,y_1),\dots,(x_p,y_p))$ be an $\mathrm{FOC}[+]$ formula holding true only when $(x_i,y_i) \models U(x,y)$ for $i = 1,\dots,p$. Then, there exists an $\mathrm{FOC}[+]$ formula

$$\psi((x,y),(x_2,y_2),\dots,(x_p,y_p))$$

which holds true for all assignments $(a_2,b_2),\dots,(a_p,b_p)$ of $(x_2,y_2),\dots,(x_p,y_p)$ exactly for the value (a,b) of (x,y), if it exists, for which

$$\rho(a,b) = |\{(a_1,b_1) \mid \phi((a_1,b_1),\cdots,(a_p,b_p))\}| \tag{2}$$

holds true.

Proof We may not directly define values which exceed `last`. So we will first explain how to manage values exceeding `last`, by representing ranks and differences of ranks less than or equal to `last`.

We explain intuitively how we proceed. Equality (1) is equivalent to

$$\rho(x_2,y_2) = \rho(x_3,y_3) - \rho(x_1,y_1). \tag{3}$$

In order to express it, we consider the graph of the total ordering enriched with a dummy element of rank 0. The value $\rho(x_3,y_3) - \rho(x_1,y_1)$ is less than $k \cdot \mathtt{last}$, which means that there exists a path of length less than or equal to k, starting from (x_1,y_1) and ending in (x_3,y_3) such that the difference of the ranks of two successive elements is less than or equal to `last`. Equality (3) can thus be expressed by stating that there exists a path of the same length starting with the element of rank 0, ending in (x_2,y_2) with the same sequence of differences of ranks between corresponding elements. We do not need this dummy element of rank 0 since the first difference is actually a rank. It thus suffices to show that we can express the predicate

$$D((u,v),(z,t)) = d \tag{4}$$

meaning that the difference of the ranks of (u,v) and (z,t) is equal to $d \leq \mathtt{last}$ and the predicate

$$R(u,v) = \rho \tag{5}$$

meaning that the rank of (u,v) is equal to $\rho \leq \mathtt{last}$. Since the first component of each element has value bounded by k, the set $\{(x,y) \mid (u,v) < (x,y) \leq (z,t)\}$ is the union, for $1 \leq i \leq k$, of

$$X_i = \{(i,y) \mid (u,v) < (i,y) \leq (z,t)\}.$$

As the definition of counting quantifier requires a nonzero value, we must select among the X_i's those which are nonempty. In order to express (4) we are thus led to consider the disjunction over all possible sequences $1 \leq i_1 < \cdots < i_r \leq k$ of the following predicates (but at most one sequence is satisfied):

$$\exists d_1 \cdots \exists d_r \Big((d = \sum_{1 \leq s \leq r} d_s) \wedge \bigwedge_{s=1}^{r} \exists_y^{d_s} ((u,v) < (i_s, y) \leq (z,t)) \Big) \wedge \\ \bigwedge_{i \neq i_1, \ldots, i_r} \neg \exists y \, ((u,v) < (i,y) \leq (z,t)).$$

The predicate (5) can be dealt with similarly:

$$\exists \rho_1 \cdots \exists \rho_r \Big((\rho = \sum_{1 \leq s \leq r} \rho_s) \wedge \bigwedge_{s=1}^{r} \exists_y^{\rho_s} ((i_s, y) \leq (u,v)) \Big) \wedge \\ \bigwedge_{i \neq i_1, \ldots, i_r} \neg \exists y \, ((i,y) \leq (u,v)).$$

Concerning the property (2), we proceed in the same manner by decomposing all pairs of elements according to the value of their first component, since this value is bounded by k. We use a formula in FOC[+] which has $2p$ variables, namely $x_2, y_2, \ldots, x_p, y_p, x, y$ and which is again a disjunction over all possible sequences $1 \leq i_1 < \cdots < i_r \leq k$ of the following formulas:

$$\exists d_1 \cdots \exists d_r \; \exists (u_1, v_1) \cdots \exists (u_r, v_r) \Big[(u_r, v_r) = (x,y) \wedge \; (R(u_1, v_1) = d_1) \wedge \\ \big(\bigwedge_{s=1}^{r-1} D((u_s, v_s), (u_{s+1}, v_{s+1})) = d_{s+1} \big) \wedge \bigwedge_{s=1}^{r} \exists_y^{d_s} \phi((i_s, y), (x_2, y_2), \ldots, (x_p, y_p)) \Big] \wedge \\ \bigwedge_{i \neq i_1, \ldots, i_r} \neg \exists y \, \phi((i,y), (x_2, y_2), \ldots, (x_p, y_p)).$$

□

We apply this result to the following property of $\mathscr{L}$(FOC[+]):

Proposition 5 *The class* $\mathscr{L}$(FOC[+]) *is closed under all inverse homomorphisms.*

Proof We use the preliminary observation of this subsection concerning the classification of morphisms. The first two cases can be treated as in Proposition 3 (counting quantifiers are straightforwardly managed). The erasing morphisms are actually considered in Proposition 8, since they can be regarded as a particular disjoint shuffle. Therefore it is enough to prove the result for growing homomorphisms. More precisely, we prove that if a language L is expressible in FOC[+] then $g^{-1}(L)$ is also expressible in FOC[+], where g is a growing morphism defined as $g(a) = ac$ and $g(b) = b$, for every $b \neq a$.

The universe of the structure $h(w)$ can be encoded into the universe $\{1, \ldots, |w|\}$ of w via the FOC[+] predicate:

$$\mathtt{U}(x,y) \equiv (y = 1 \vee (Q_a(x) \wedge y = 2)).$$

The rank $\rho(i,j)$ of a pair (i,j) is defined according to the alphabetic ordering and satisfies the condition

$$\rho(i,j) = i + |\{x \mid Q_a(x) \wedge x < i\}| + j - 1. \tag{6}$$

Given a formula $\phi(x_1,\ldots,x_n)$ in FOC[+] and a word $h(w)$, we show how to associate a formula $\phi'((x_1,y_1),\ldots,(x_n,y_n))$ in FOC[+] such that the following holds:

$$\begin{gathered}(w,\rho^{-1}(i_1),\ldots,\rho^{-1}(i_n)) \models \phi'((x_1,y_1),\ldots,(x_n,y_n))\\ \Leftrightarrow\\ (h(w),i_1,\ldots,i_n) \models \phi(x_1,\ldots,x_n).\end{gathered}$$

We proceed by structural induction starting with the basic predicates.

If $\phi(x) \equiv Q_b(x)$ then

$$\phi'(x,y) \equiv \begin{cases} Q_a(x) \wedge y = 1 & \text{if } b = a \\ Q_a(x) \wedge y = 2 & \text{if } b = c \\ Q_b(x) \wedge y = 1 & \text{otherwise,} \end{cases}$$

and the predicate $x = y+z$ is replaced by the predicate of Proposition 4(i).

With the formula $\neg\phi(x_1,\ldots,x_n)$, we associate the formula

$$\neg\phi'((x_1,y_1),\ldots,(x_n,y_n)) \wedge \bigwedge_{1\leq i\leq n} \mathtt{U}(x_i,y_i).$$

Concerning the conjunction, assume without loss of generality that the two formulas are $\phi_1(x_1,\ldots,x_n,\ldots,x_{n+m})$ and $\phi_2(x_{n+1},\ldots,x_{n+m},\ldots x_{n+m+p})$, i.e., that the common variables occur among the last and first ones in the formulas, respectively. Then, with the formula

$$\phi_1(x_1,\ldots,x_{n+m}) \wedge \phi_2(x_{n+1},\ldots,x_{n+m+p})$$

we associate the formula

$$\phi_1'((x_1,y_1),\ldots,(x_{n+m},y_{n+m})) \wedge \phi_2'((x_{n+1},y_{n+1}),\ldots,(x_{n+m+p},y_{n+m+p})).$$

Finally, the rule for the counting quantifiers is given in Proposition 4(ii). □

4.2 Concatenation, Shuffle and Kleene Star

Let us start by considering the closure under concatenation:

Proposition 6 *The classes* $\mathscr{L}(\mathrm{FO}[<])$, $\mathscr{L}(\mathrm{FO}[+])$ *and* $\mathscr{L}(\mathrm{FOC}[+])$ *are closed under concatenation, while the class* $\mathscr{L}(\mathrm{FO}[+1])$ *is not.*

Proof As a preliminary observation, we assume that in the formulas ϕ_1 and ϕ_2 defining two languages L_1 and L_2, respectively, all quantified variables are expressed in the form $\exists x \leq \mathtt{last}$ and $\forall x \leq \mathtt{last}$ and that no variable, whether free or bound, is shared by both formulas. We apply the method of relativization: if $u \in L_1$ and $v \in L_2$ then uv is defined by saying that u satisfies ϕ_1, and that the "translate" of the factor v obtained by shifting it $|u|$ positions to the left satisfies ϕ_2. The new formula is therefore of the form

$$\psi = \exists\ell\ (\phi_1' \wedge \phi_2').$$

The case $\mathscr{L}(\mathrm{FO}[<])$ can be treated directly, but the result is a simple consequence of the fact that it is a pseudovariety of languages.

Concerning $\mathscr{L}(\mathrm{FO}[+])$, we apply two different sets of transformations on the formulas with arbitrary free variables according to whether we define the prefix or the suffix of the word $w = uv$:

- ϕ_1' is obtained from ϕ_1 by replacing `last` by ℓ, every occurrence of $Q_a(x)$ by $Q_a(x) \wedge x \leq \ell$, every occurrence of $z = x + y$ by $z = x + y \wedge (x \leq \ell) \wedge (y \leq \ell) \wedge (z \leq \ell)$. Straightforward substitutions can manage the inductive construction of formulas.
- ϕ_2' is obtained from ϕ_2 by replacing every occurrence of $Q_a(x)$ by $Q_a(x) \wedge x > \ell$, every occurrence of $z = x + y$ by $z = x + y - \ell \wedge (x > \ell) \wedge (y > \ell) \wedge (z > \ell)$. Straightforward substitutions can manage the inductive construction of formulas.

It suffices to check by structural induction that the following holds

$$\begin{gathered}(u, i_1, \ldots, i_r) \models \phi_1 \text{ and } (v, j_1, \ldots, j_s) \models \phi_2 \\ \Leftrightarrow \\ (uv, i_1, \ldots, i_r, j_1 + |u|, \ldots, j_s + |u|) \models \phi_1' \wedge \phi_2'.\end{gathered} \tag{7}$$

We now deal with the third language class, and assume ϕ_1 and ϕ_2 are FOC[+] formulas:

- ϕ_1' is obtained from ϕ_1 as above with the additional rule
  - $\exists^y_{x_1}\, \theta(x_1, \ldots, x_n)$ replaced by $\exists^y_{x_1}\, (\theta(x_1, \ldots, x_n) \wedge (x_1 \leq \ell))$.
- ϕ_2' is obtained from ϕ_2 as above with the additional rule
  - $\exists^y_{x_1}\, \theta(x_1, \ldots, x_n)$ replaced by $\exists t\, (\exists^t_{x_1}\, \theta(x_1, \ldots, x_n) \wedge (x_1 > \ell)) \wedge (y = t + \ell))$.

  where t is a new variable. With these two modifications, it suffices to verify a formula as (7).

For the last statement, it suffices to observe that the languages c^*ac^* and c^*bc^* are in $\mathscr{L}(\mathrm{FOC}[+1])$ but not their concatenation (see, e.g., [18, Cor. IV.3.4]). □

We now turn to the shuffle operation which assigns a finite subset of words to a pair of words by the following recursive definition:

$$\begin{aligned}& w \sqcup\!\sqcup\, \varepsilon = \varepsilon \sqcup\!\sqcup\, w = w, \\ & au \sqcup\!\sqcup\, bv = b(au \sqcup\!\sqcup\, v) \cup a(u \sqcup\!\sqcup\, bv), \text{with } a, b \in \Sigma \text{ and } u, v \in \Sigma^*.\end{aligned}$$

E.g., if $u = aba$ and $v = aa$ then $u \sqcup\!\sqcup\, v = \{a^3ba, a^2ba^2, aba^3\}$. This notation extends to languages $A, B \subseteq \Sigma^*$ by defining $A \sqcup\!\sqcup\, B = \bigcup_{x \in A, y \in B} x \sqcup\!\sqcup\, y$. A *disjoint shuffle* is a shuffle performed between two languages defined over disjoint alphabets.

Proposition 7 *The classes $\mathscr{L}(\mathrm{FO}[+1])$, $\mathscr{L}(\mathrm{FO}[<])$ and $\mathscr{L}(\mathrm{FO}[+])$ are not closed under shuffle.*

Proof This assertion is proved via counterexamples. The languages $X = c^*$ and $Y = \{ab\}$ are in $\mathscr{L}(\mathrm{FO}[+1])$ but $X \sqcup\!\sqcup\, Y$ is not. The languages $X = b^*$, $Y = (a^2b)^*$ and $Z = (ab)^*$ are in $\mathscr{L}(\mathrm{FO}[<])$ but $(X \sqcup\!\sqcup\, Y) \cap Z = ((ab)^2)^*$ is not. Finally, the languages $X = (a^2)^*$ and $Y = b^*$ belong to $\mathscr{L}(\mathrm{FO}[+])$ but $X \sqcup\!\sqcup\, Y$ does not by [2, Cor. 4.2]. □

The situation changes for disjoint shuffles:

Proposition 8 *The classes $\mathscr{L}(\mathrm{FO}[<])$ and $\mathscr{L}(\mathrm{FOC}[+])$ are closed under disjoint shuffle, while the classes $\mathscr{L}(\mathrm{FO}[+1])$ and $\mathscr{L}(\mathrm{FO}[+])$ are not.*

Proof Let the formulas ϕ_1 and ϕ_2 define the languages $L_1 \subseteq \Sigma_1^*$ and $L_2 \subseteq \Sigma_2^*$, respectively, with $\Sigma_1 \cap \Sigma_2 = \emptyset$. Given $u \in \Sigma_1^*$ and $v \in \Sigma_2^*$, let $w = w_1 \cdots w_{r+s} \in u \sqcup\!\sqcup v$ with $|u| = r$ and $|v| = s$. For $i = 1, \ldots r$, let $f(i)$ be the position in w of the ith letter of u, and similarly let $g(j)$ be the position in w of the jth letter of v. Example: with $u = aba$, $v = ddcd$ and $w = ddabcda$ we have $f(1) = 3$, $f(2) = 4$, $f(3) = 7$ and $g(1) = 1$, $g(2) = 2$, $g(3) = 5$, $g(4) = 6$.

The statement for FO[<] comes straightforwardly since the syntactic monoid of $L_1 \sqcup\!\sqcup L_2$ is the direct product of the syntactic monoids of L_1 and L_2. Since each of these two syntactic monoids has trivial groups only, so does the direct product.

Concerning FOC[+], we assume without loss of generality that the formulas ϕ_1 and ϕ_2 do not contain standard existential quantifiers since they can be simulated by counting quantifiers, that the symbol `last` does not occur and that the two sets of their variables, whether free or bound, are disjoint. In order to simplify the notations, we let $Q_{\Sigma_i}(x)$ stand for $\bigvee_{a \in \Sigma_i} Q_a(x)$. We denote by ϕ' the formula resulting from ϕ and we drop the index since for obvious reasons of symmetry, the two formulas are modified according to the same rules. We proceed according to the usual structural induction, by applying the following rules:

- If $\phi(x) \equiv Q_a(x)$ then $\phi'(x) \equiv Q_a(x)$.
- If $\phi(x,y,z) \equiv x = y + z$ then

$$\phi'(x,y,z) \equiv \exists x' \exists y' \exists z' \Big((x' = y' + z') \wedge \exists_u^{y'} (Q_\Sigma(u) \wedge u \leq y) \wedge \\ \exists_u^{z'} (Q_\Sigma(u) \wedge u \leq z) \wedge \exists_u^{x'} (Q_\Sigma(u) \wedge u \leq x) \Big).$$

- If $\phi \equiv \neg \psi(x_1, \ldots, x_n)$ then

$$\phi' \equiv \neg \psi'(x_1, \ldots, x_n) \wedge \bigwedge_{i=1}^{n} Q_\Sigma(x_i).$$

- If $\phi \equiv \psi \wedge \theta$ then $\phi' \equiv \psi' \wedge \theta'$.
- If $\phi(x) \equiv \exists_u^x \psi(u)$ then $\phi'(x) \equiv \exists_u^x (\psi'(u) \wedge Q_\Sigma(u))$.

It suffices to verify by structural induction that if ϕ_1' and ϕ_2' are the formulas obtained from ϕ_1 and ϕ_2 by applying the above rules, then we have

$$(u, i_1, \ldots, i_r) \models \phi_1(x_1, \ldots, x_r) \text{ and } (v, j_1, \ldots, j_s) \models \phi_2(y_1, \ldots, y_s) \\ \Leftrightarrow \\ (w, f(i_1), \ldots, f(i_r), g(j_1), \ldots, g(j_s)) \models \phi_1'(x_1, \ldots, x_r) \wedge \phi_2'(y_1, \ldots, y_s).$$

To see that $\mathscr{L}(\mathrm{FO}[+1])$ and $\mathscr{L}(\mathrm{FO}[+])$ are not closed under disjoint shuffle, it is enough to check counterexamples provided in Proposition 7 for general shuffle. □

Finally, let us consider the closure under Kleene star:

Proposition 9 *The classes* $\mathscr{L}(\mathrm{FO}[+1])$*,* $\mathscr{L}(\mathrm{FO}[<])$ *and* $\mathscr{L}(\mathrm{FO}[+])$ *are not closed under Kleene star.*

Proof The counterexample for $\mathscr{L}(\mathrm{FO}[+1])$ and $\mathscr{L}(\mathrm{FO}[<])$ is the subset $(a^2)^*$. Now, if $\mathscr{L}(\mathrm{FO}[+])$ were closed under Kleene star, by Proposition 6 the whole class of regular languages would be contained in $\mathscr{L}(\mathrm{FO}[+])$ due to Kleene's Theorem. However, the regular language $\{w \in \{a,b\}^+ \mid |w|_a \bmod 2 = 0\}$ does not belong to $\mathscr{L}(\mathrm{FO}[+])$ by [2, Cor. 4.2], a contradiction. □

To the best of our knowledge, the status of $\mathscr{L}(\mathrm{FOC}[+])$ relative to the Kleene star is still open. All we can say is that:

If $\mathscr{L}(\mathrm{FOC}[+])$ were closed under Kleene star, then $\mathrm{TC}^0 = \mathrm{NC}^1$.

Indeed, in that case $\mathscr{L}(\mathrm{FOC}[+]) \subset \mathrm{TC}^0$ would be closed under the regular operations and thus it would contain all regular languages. However, as proved in [3], there exists a regular language which is complete for NC^1.

4.3 Quotient

The *left quotient* of a language L by a letter a, denoted $a^{-1}L$, is the set of words u such that $au \in L$. In this subsection, "quotient" means left quotient. For obvious reasons of symmetry, the same result holds for right quotients.

Proposition 10 *The classes $\mathscr{L}(\mathrm{FO}[+1])$, $\mathscr{L}(\mathrm{FO}[<])$, $\mathscr{L}(\mathrm{FO}[+])$, $\mathscr{L}(\mathrm{FOC}[+])$ are closed under quotient.*

Proof We encode the universe of aw into the universe of w as the set of pairs (x,y) which satisfy the predicate

$$\mathtt{U}(x,y) \equiv ((x=1 \wedge y=1) \vee x=2).$$

The rank of an element (i,j) of this set relative to the alphabetic ordering is $\rho(i,j)=i+j-1$, and can be defined in the logics FO[+] and FOC[+]. We encode the predicates Q_b and $+1$, of the universe of aw into the universe of w as follows, respectively:

$$P_b(x,y) \equiv \begin{cases} (x=1 \wedge y=1) \vee (x=2 \wedge Q_a(y)) & \text{if } b=a \\ x=2 \wedge Q_b(y) & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \mathtt{Succ}((x_1,y_1),(x_2,y_2)) \equiv\ & (x_1=y_1=y_2=1 \wedge x_2=2) \vee \\ & ((x_1=x_2=2) \wedge (y_2=y_1+1)). \end{aligned}$$

If $\phi \in \mathrm{FO}[+1]$ defines the language L then the language $a^{-1}L$ is defined by the formula ϕ' which is obtained by recursively applying the following rules:

- If $\phi(x) \equiv Q_b(x)$ then $\phi'(x,y) \equiv P_b(x,y)$.
- If $\phi(x_1,x_2) \equiv x_2 = x_1+1$ then $\phi'((x_1,y_1),(x_2,y_2)) \equiv \mathtt{Succ}((x_1,y_1),(x_2,y_2))$.
- If $\psi(x_1,\dots,x_n) \equiv \neg\phi(x_1,\dots,x_n)$ then

$$\psi'((x_1,y_1),\dots,(x_n,y_n)) \equiv \neg\phi'((x_1,y_1),\dots,(x_n,y_n)) \wedge \bigwedge_{i=1}^{n} \mathtt{U}(x_i,y_i).$$

- If $\psi \equiv \phi_1 \wedge \phi_2$ then $\psi' \equiv \phi'_1 \wedge \phi'_2$.
- If $\psi(x_1,x_2,\dots,x_n) \equiv \exists x_1\, \phi(x_1,x_2,\dots,x_n)$ then

$$\psi'((x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)) \equiv \exists x_1 \exists y_1\, \phi'((x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)) \wedge \mathtt{U}(x_1,y_1).$$

The assertion concerning FO[<] follows from the fact that $\mathscr{L}(\mathrm{FO}[<])$ is a pseudovariety. In order to treat the case FO[+], we encode the predicate $x+y=z$ by the predicate which asserts that the rank of (z_1,z_2) in the alphabetic ordering is the sum of the ranks of (x_1,x_2) and (y_1,y_2), i.e., $\rho(x_1,y_1)+\rho(x_2,y_2)=\rho(x_3,y_3)$, which is clearly expressible in FO[+]. Then, we have

$$(w,(i_1,j_1),\ldots,(i_n,j_n)) \models \phi'((x_1,y_1),\ldots,(x_n,y_n))$$
$$\Leftrightarrow$$
$$(aw,\rho(i_1,j_1),\ldots,\rho(i_n,j_n)) \models \phi(x_1,\ldots,x_n).$$

For FOC[+], the only new transformation rule is the one involving $\exists_{x_1}^{y}\phi(x_1,\ldots,x_n)$. By induction, with the formula ϕ for L the formula $\phi'((x_1,y_1),\ldots,(x_n,y_n))$ for $a^{-1}L$ is assigned. Indeed, we replace the formula $\exists_{x_1}^{y}\phi(x_1,\ldots,x_n)$ with the formula

$$\psi((z_1,t_1),(x_2,y_2),\ldots,(x_n,y_n))$$

whose truth value for a fixed assignment $(a_2,b_2),\ldots,(a_n,b_n)$ can be described as follows: Let r be the number of values (a_1,b_1) such that $\phi'((a_1,b_1),(a_2,b_2),\ldots,(a_n,b_n))$ holds true. If $r=0$, $\psi((z_1,t_1),(a_2,b_2),\ldots,(a_n,b_n))$ is a contradiction, otherwise it assumes the value `true` for the unique pair (a_1,b_1) such that $\rho(a_1,b_1)=r$. The implementation of ψ is given in Proposition 4(ii). □

4.4 Conjugacy

We recall that two words x and y are *conjugate* if there exist u and v such that $x=uv$ and $y=vu$. More generally, the conjugate of a language L is the set of all conjugates of the words it contains, i.e., it is the set $\mathrm{CONJ}(L)=\{vu\in\Sigma^*\mid uv\in L\}$.

Proposition 11 *The classes* $\mathscr{L}(\mathrm{FO}[+1])$, $\mathscr{L}(\mathrm{FO}[<])$, $\mathscr{L}(\mathrm{FO}[+])$, $\mathscr{L}(\mathrm{FOC}[+])$ *are closed under conjugacy.*

Proof A language is expressible in FO[+1] if and only if it is *locally threshold testable* [18], i.e., for some integers k and T it is a union of classes of an equivalence relation $u\sim v$ defined by the following conditions:

- u and v have the same prefix and the same suffix of length $k-1$,
- for all w of length k, either $|u|_w=|v|_w<T$ or $|u|_w,|v|_w\geq T$ holds.

It thus suffices to consider a single equivalence class consisting of words of length greater than k, i.e. to fix $k,T\in\mathbb{N}$ and $p,s\in\Sigma^{k-1}$, a mapping $f:\Sigma^k\to\{0,\ldots,T\}$, and to consider the language of the words u satisfying

$$u\in p\Sigma^*\cap\Sigma^*s \text{ and, for all } w\in\Sigma^k,\ |u|_w=f(w) \text{ if } f(w)<T \text{ or else } |u|_w\geq T.$$

Let us introduce the following definition. A word w is a *cyclic occurrence* of u if $u=xwy$ holds or if there exists a factorization $w=w_1w_2$, with $w_1,w_2\neq\varepsilon$, and a word x such that $u=w_2xw_1$ holds. The cyclic occurrence *starts* (resp., *ends*) *at position* k if $|x|=k$ in the first case and if $|w_2x|=k$ in the second case (resp., if $|xw|=k$ in the first case and if $|w_2|=k$ in the second case). Moreover, it *avoids* position ℓ if it does not contain the ℓth letter of u. Then, the set of conjugates of this equivalence class is defined by a formula which says that for some integer ℓ, the word has a cyclic occurrence of p starting at position ℓ, a cyclic occurrence of s ending at position ℓ and such that for all words w of length k, it has exactly

$f(w)$ cyclic occurrences w avoiding ℓ if $f(w) < T$ or at least T cyclic occurrences avoiding ℓ otherwise.

The result concerning FO[$<$] is well-known. We briefly recall it for the sake of self containment. Consider a deterministic aperiodic automaton[1] recognizing the language L, with q_0 as initial state and F as set of final states. Then the set of conjugates of L is the union of the languages Y_qZ_q, where Y_q is the set of words recognized by the automaton with q as initial state and F as set of final state and Z_q is the set of words recognized by the automaton with q_0 as initial state and where the set of final states reduces to q. These automata are aperiodic, thus they recognize star-free languages. We conclude by recalling that, by definition, star-free languages are closed under union and concatenation.

For the logic FO[$+$], we proceed as follows. Given a formula ϕ defining a language L, we construct a formula $\exists t\, \phi'(t)$ defining the conjugate of L. The variable t under the existential quantifier must be interpreted as the shift of a word in L. Indeed, consider the following function describing, for all fixed values of the variable t, a bijection of $\{1,\dots,\texttt{last}\}$ into itself which defines where the position of a letter in uv maps in vu:

$$f(x,t) = \begin{cases} x + \texttt{last} - t & \text{if } x \le t \\ x - t & \text{otherwise.} \end{cases} \tag{8}$$

Observe that this correspondence is definable in FO[$+$].

From a formula $\phi(x_1,\dots,x_n)$ in FO[$+$], we define a formula $\phi'(x_1,\dots,x_n,t)$ with the property

$$\begin{gathered} (uv, i_1,\dots,i_n) \models \phi(x_1,\dots,x_n) \\ \Leftrightarrow \\ (vu, f(i_1,|u|),\dots,f(i_n,|u|),|u|) \models \phi'(x_1,\dots,x_n,t) \end{gathered} \tag{9}$$

by applying the usual structural induction (we assume the variable t does not occur in the formula ϕ):

- If $\phi(x) \equiv Q_a(x)$ then $\phi'(x,t) \equiv Q_a(f(x,t))$.
- If $\phi(x,y,z) \equiv z = x + y$ then $\phi'(x,y,z,t) \equiv f(z,t) = f(x,t) + f(y,t)$.
- If $\phi(x_1,\dots,x_n) \equiv \phi_1(y_1,\dots,y_p) \wedge \phi_2(z_1,\dots,z_m)$, where $x_1,\dots,x_n$ is the set consisting of the variables of ϕ_1 and of ϕ_2, then $\phi'(x_1,\dots,x_n,t) \equiv \phi_1'(y_1,\dots,y_p,t) \wedge \phi_2'(z_1,\dots,z_m,t)$.
- If $\phi \equiv \neg\psi$ then $\phi' \equiv \neg\psi'$.
- If $\phi \equiv \exists x_1\, \psi(x_1,x_2,\dots,x_n)$ then $\phi' \equiv \exists x_1\, \psi'(x_1,x_2,\dots,x_n)$.

So, if L is defined by the closed formula ϕ then its conjugate is defined by the formula $\exists t\, \phi'(t)$.

We now turn to the logic FOC[$+$]. It suffices to show how to deal with the counting quantifier $\exists^{y}_{x_1}\, \phi(x_1,\dots,x_n)$. This expression is recursively transformed into

$$\begin{aligned} &\exists y_1 \exists y_2 \left(f(y,t) = f(y_1,t) + f(y_2,t)\right) \wedge \\ &\quad \exists^{y_1}_{x_1} \left(\phi'(x_1,\dots,x_n,t) \wedge (x_1 \le \texttt{last} - t)\right) \wedge \\ &\quad \exists^{y_2}_{x_1} \left(\phi'(x_1,\dots,x_n,t) \wedge (x_1 > \texttt{last} - t)\right). \end{aligned}$$

Similarly to FO[$+$], if a language L is defined by a closed formula ϕ then its conjugate is defined by the formula $\exists t\, \phi'(t)$. □

[1] A deterministic automaton is aperiodic if and only if there exists $n > 0$ such that $q \cdot w^n = q \cdot w^{n+1}$, for any state q and any word w.

4.5 Commutative Image

The *commutative image* of a language L is the language

$$\mathrm{COMM}(L) = \{x_1 \cdots x_n \in \Sigma^* \mid x_{i_1} \cdots x_{i_n} \in L \text{ and } \{i_1, \ldots, i_n\} = \{1, \ldots, n\}\}.$$

Proposition 12 *The classes $\mathscr{L}(\mathrm{FO}[+1])$, $\mathscr{L}(\mathrm{FO}[<])$ and $\mathscr{L}(\mathrm{FO}[+])$ are not closed under commutative image.*

Proof For $\mathscr{L}(\mathrm{FO}[+1])$ and $\mathscr{L}(\mathrm{FO}[<])$, it suffices to see that $\mathrm{COMM}((ab)^*)$ is not a regular language. Now, let $L = \{abab^2ab^3a \cdots ab^ia \cdots ab^ka \mid k > 0\}$. By the proof of Proposition 2, we have that $L \in \mathrm{FO}[+]$. Consider the language $L' = \mathrm{COMM}(L) \cap a^*b^* = \{a^n b^{\frac{(n-1)n}{2}} \mid n > 1\}$. If $\mathscr{L}(\mathrm{FO}[+])$ were closed under commutative image, then L' should be semilinear by Theorem 2, a contradiction. □

4.6 Reversal

Given a word $x = x_1 \cdots x_n$, with $x_i \in \Sigma$, its reversal is the word $x^R = x_n \cdots x_1$ (with $\varepsilon^R = \varepsilon$). The *reversal* of a language L is the language $L^R = \{x^R \mid x \in L\}$.

Proposition 13 *The classes $\mathscr{L}(\mathrm{FO}[+1])$, $\mathscr{L}(\mathrm{FO}[<])$, $\mathscr{L}(\mathrm{FO}[+])$, $\mathscr{L}(\mathrm{FOC}[+])$ are closed under reversal.*

Proof Let ϕ be a formula defining a language in $\mathscr{L}(\mathrm{FO}[+1])$, $\mathscr{L}(\mathrm{FO}[<])$, $\mathscr{L}(\mathrm{FO}[+])$ and $\mathscr{L}(\mathrm{FOC}[+])$, respectively. An FO[+1] formula for L^R is obtained from ϕ by replacing each occurrence of $x = y + 1$ by $y = x + 1$. Concerning $\mathscr{L}(\mathrm{FO}[<])$, it suffices to replace $x < y$ by $y < x$. Concerning $\mathscr{L}(\mathrm{FO}[+])$, we substitute $x + y - \mathtt{last} - 1 = z$ for each occurrence of $x + y = z$ and, in addition for FOC[+], we substitute $\exists^z_{x_1} \phi(x_1, \ldots, x_p) \wedge (z + y = \mathtt{last} + 1)$ for $\exists^y_{x_1} \phi(x_1, \ldots, x_p)$. □

5 An Application: Dyck Languages

From [1], we know that the Dyck languages are in $\mathscr{L}(\mathrm{FOC}[+])$, while from [14] we get that they do not belong to $\mathscr{L}(\mathrm{FO}[+])$. As an application of closure properties in the previous section, we are now going to provide an alternative proof of this latter fact.

The following preliminary notions are useful. Let A be a finite set of opening parentheses and let $\overline{A}$ be the set of (one-to-one) corresponding closing parentheses. Set $T = A \cup \overline{A}$; a word in T^* is correctly (or well) parenthesized if: (i) any opening parenthesis a is followed by a corresponding closing parenthesis $\overline{a}$, and (ii) if parenthesis a' follows a, then a' is closed before a. The *Dyck language* D_T is the set of correctly parenthesized words in T^*.

The *majority* function $\mathscr{M}_{\sigma,\overline{\sigma}} : \{\sigma, \overline{\sigma}\}^* \rightarrow \{0,1\}$ and the *equality* function $\mathscr{E}_{\sigma,\overline{\sigma}} : \{\sigma, \overline{\sigma}\}^* \rightarrow \{0,1\}$ are defined, respectively, as:

$$\mathscr{M}_{\sigma,\overline{\sigma}}(x) = \begin{cases} 1 & \text{if } |x|_\sigma > |x|_{\overline{\sigma}} \\ 0 & \text{otherwise,} \end{cases} \qquad \mathscr{E}_{\sigma,\overline{\sigma}}(x) = \begin{cases} 1 & \text{if } |x|_\sigma = |x|_{\overline{\sigma}} \\ 0 & \text{otherwise.} \end{cases}$$

We also need the notion of AC^0-reduction between problems [18,19]. Informally, a problem P is AC^0-reducible to a problem P' whenever P can be solved by a family of polynomial size, constant depth, unbounded fan-in AND/OR/NOT-circuits with *oracle gates* for P'. In this case, it is easy to see that $P' \in \mathrm{AC}^0$ implies $P \in \mathrm{AC}^0$ as well.

Theorem 5 *The Dyck language $D_{\{a,\overline{a}\}}$ does not belong to $\mathscr{L}(\mathrm{FO}[+])$.*

Proof We first prove that $\mathscr{E}_{a,\overline{a}}$ is not in AC^0. Indeed, since $\mathscr{M}_{a,\overline{a}}$ is not in AC^0 (see, e.g., [7]), it suffices to show that $\mathscr{M}_{a,\overline{a}}$ is AC^0-reducible to $\mathscr{E}_{a,\overline{a}}$. Consider $x = x_1 \cdots x_n \in \{a,\overline{a}\}^n$. To compute $\mathscr{M}_{a,\overline{a}}(x)$, we build an AC^0-circuit C_n containing a first layer of oracle gates $O_0, \ldots, O_{\lfloor \frac{n}{2} \rfloor}$ for $\mathscr{E}_{a,\overline{a}}$. As input to O_i, we give the word $w(i) = a^{i+(n \bmod 2)} x_{i+1} \cdots x_n$. If there is an oracle O_i yielding 1, we have that $|x|_a \leq |w(i)|_a = |w(i)|_{\overline{a}} \leq |x|_{\overline{a}}$. On the contrary, if all O_i's yield 0, we get that $|x|_a > |x|_{\overline{a}}$. Thus, we can complete C_n by plugging all the outputs of O_i's into an OR gate whose output, in turn, is sent to a final NOT gate.

Let us now turn to the Dyck language. First of all, it is not difficult to verify that

$$\mathrm{CONJ}(D_{\{a,\overline{a}\}}) = \{w \in \{a,\overline{a}\}^* \mid |w|_a = |w|_{\overline{a}}\}.$$

Now, assume by contradiction that $D_{\{a,\overline{a}\}} \in \mathscr{L}(\mathrm{FO}[+])$. Proposition 11 says that $\mathscr{L}(\mathrm{FO}[+])$ is closed under conjugation, thus implying that $\mathrm{CONJ}(D_{\{a,\overline{a}\}}) \in \mathscr{L}(\mathrm{FO}[+])$ as well. However, we have $w \in \mathrm{CONJ}(D_{\{a,\overline{a}\}})$ if and only if $\mathscr{E}_{a,\overline{a}}(w) = 1$. Since $\mathscr{E}_{a,\overline{a}}$ is not in AC^0 and since $\mathscr{L}(\mathrm{FO}[+]) \subset \mathrm{AC}^0$, we get the result. □

Theorem 6 *The Dyck language D_T does not belong to $\mathscr{L}(\mathrm{FO}[+])$.*

Proof Let $T = A \cup \overline{A}$ and $a \in A$ a type of parentheses of D_T. By contradiction, suppose there exists an FO[+] formula ϕ for D_T, and construct the formula $\phi' = \phi \wedge \forall x\,(Q_a(x) \vee Q_{\overline{a}}(x))$. Clearly, ϕ' is an FO[+] formula for the subset of D_T consisting of the well parenthesized words over the alphabet $\{a,\overline{a}\}$, namely $D_{\{a,\overline{a}\}}$. This contradicts Theorem 5. □

References

1. D.A. Mix Barrington and J. Corbett. On the relative complexity of some languages in NC. *Information Processing Letters*, 32:251–256, 1989.
2. D.A. Mix Barrington, N. Immerman, C. Lautemann, N. Schweikardt and D. Thérien. First-order expressibility of languages with neutral letters or: The Crane Beach conjecture. *J. Comput. Syst. Sci.*, 70:101–127, 2005.
3. D.A. Mix Barrington, K.J. Compton, H. Straubing and D Thérien. Regular languages in NC^1. *J. Comput. Syst. Sci.*, 44(3):478–499, 1992.
4. J.R. Büchi. Weak second order arithmetic and finite automata. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 6:66–92, 1960.
5. N. Chomsky and M.P. Schützenberger. The algebraic theory of context-free languages. In P. Braffort and D. Hirschberg, Eds., *Computer Programming and Formal Systems*, pp. 118–161. North Holland, 1963.
6. K. Etessami. Counting quantifiers, successor relations, and logarithmic space. *J. Comput. Syst. Sci.*, 54(3):400–411, 1997.
7. M. Furst, J.B. Saxe and M. Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical Systems Theory*, 17:13–27, 1984.
8. S. Ginsburg. *The Mathematical Theory of Context-Free Languages*. McGraw-Hill, 1966.
9. S. Ginsburg and E.H. Spanier. Bounded ALGOL-like languages. *Trans. Amer. Math. Soc.*, 113:333–368, 1964.
10. M.A. Harrison. *Introduction to Formal Languages*. Addison-Wesley, 1978.
11. O. Ibarra. Simple matrix grammars. *Information and Control*, 17:359–394, 1970.
12. O. Ibarra. A note on semilinear sets and bounded-reversal multihead pushdown automata. *Information Processing Letters*, 3:25–28, 1974.
13. O. Ibarra, T. Jiang and B. Ravikumar. Some subclasses of context-free languages in NC^1. *Information Processing Letters*, 29:111–117, 1988.
14. D. Robinson. *Parallel algorithms for group word problems.* Doctoral Dissertation, Mathematics Dept., University of California, San Diego, 1993.

15. V. Roychowdhury, K.-Y. Siu and A. Orlitsky. Neural models and spectral methods. In: V. Roychowdhury, K.-Y. Siu, and A. Orlitsky, Eds., *Theoretical advances in Neural Computation and Learning*, pp. 3–36. Kluwer Academic, 1994.
16. W.L. Ruzzo. On uniform circuit complexity. *J. Comput. Syst. Sci.*, 22:365–383, 1981.
17. N. Schweikardt. Arithmetic, first-order logic, and counting quantifiers. *ACM Trans. Comput. Log.*, 6(3):634–671, 2005.
18. H. Straubing. *Finite Automata, Formal Logic, and Circuit Complexity*. Birkhäuser, 1994.
19. I. Wegener. *The Complexity of Boolean Functions*. Teubner, 1987.