Languages ordered by the subword order*

Dietrich Kuske[1] and Georg Zetzsche[2]

[1] Technische Universität Ilmenau, Germany
dietrich.kuske@tu-ilmenau.de
[2] Max Planck Institute for Software Systems (MPI-SWS), Germany
georg@mpi-sws.org

Abstract. We consider a language together with the subword relation, the cover relation, and regular predicates. For such structures, we consider the extension of first-order logic by threshold- and modulo-counting quantifiers. Depending on the language, the used predicates, and the fragment of the logic, we determine four new combinations that yield decidable theories. These results extend earlier ones where only the language of all words without the cover relation and fragments of first-order logic were considered.

Keywords: Subword order · First-order logic · Counting quantifiers · Decidable theories

1 Introduction

The subword relation (sometimes called scattered subword relation) is a simple example of a well-quasi ordering [7]. This property allows its prominent use in the verification of infinite-state systems [4]. The subword relation can be understood as embeddability of one word into another. This embeddability relation has been considered for other classes of structures like trees, posets, semilattices, lattices, graphs etc. [13, 15, 14, 8–11, 21, 22].

We are interested in logics over the subword order. Prior work on this has concentrated on first-order logic where the universe consists of all words over some alphabet. In this setting, we already have a rather precise picture about the border between decidability and undecidability: For the subword order alone, the $\exists^*$-theory is decidable [16] and the $\exists^*\forall^*$-theory is undecidable [12, 6]. If we add constants to the signature, already the $\exists^*$-theory becomes undecidable [6]. With regular predicates, the two-variable theory is decidable, but the three-variable theory is undecidable [12].

Thus, the decidable theories identified so far leave little room to express natural properties. First, the universe is confined to the set of all words and predicates for subsets quickly incur undecidability. Moreover, neither in the $\exists^*$-,

* Part of the results were obtained when the second author was affiliated with LSV (ENS Paris-Saclay) and supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD) and by Labex DigiCosme, Université Paris-Saclay, project VERICONISS.

nor in the two-variable fragment of first-order logic, one can express the cover relation $\lessdot$ (i.e., "u is a proper subword of v and there is no word properly between these two"). As another example, one cannot express threshold properties like "there are at most k subwords with a given property" in any of these two logics.

In this paper, we aim to identify decidable logics that are more expressive. To that end, we consider four additions to the expressivity of the logic:

- Instead of all words over some alphabet, the universe is a language L.
- We add regular predicates or constants to the structure.
- Besides the subword order, we also consider the cover relation $\lessdot$.
- We add threshold and modulo counting quantifiers to the logic.

Formally, this means we consider structures of the form

$$(L, \sqsubseteq, \lessdot, (K \cap L)_{K \text{ regular}}, (w)_{w \in L}),$$

where the universe is a language $L \subseteq \Sigma^*$, $\sqsubseteq$ is the subword ordering, $\lessdot$ is the cover relation, there is a predicate $K \cap L$ for each regular $K \subseteq \Sigma^*$, and a constant symbol for each $w \in L$. Moreover, we consider fragments of the logic C+MOD, which extends first-order logic by threshold- and modulo-counting quantifiers.

The key idea of this paper is to find decidable theories by varying the universe L and thereby either (i) simplify the structure $(L, \sqsubseteq)$ enough to obtain decidability even with the extensions above or (ii) generalize existing results that currently only apply to $L = \Sigma^*$. This leads to the following results.

1. First, we require L to be bounded. This means, we have $L \subseteq w_1^* \cdots w_m^*$ for some words $w_1, \ldots, w_m \in \Sigma^*$. Then, as soon as L is context-free, the C+MOD-theory of the whole structure is decidable (Theorem 3.4).
2. To lift the boundedness restriction, we show that if L is regular, we still obtain decidability for the whole structure if we stay within the two-variable fragment C+MOD^2 (Corollary 4.8). This generalizes the decidability of the FO^2-theory without the cover relation as shown in [12, Theorem 5.5].
3. Moreover, we consider a regular universe, but lift the two-variable requirement. To get decidability, we restrict quantifiers and available predicates: We show that for regular L, the Σ_1-theory of the structure $(L, \sqsubseteq)$ is decidable (Theorem 5.1). In the case $L = \Sigma^*$, this had been shown in [16, Prop. 2.2].
4. Finally, we place a further restriction on L, but in return obtain decidability with constants. We show that if L is regular and every letter is "frequent" in L (see section 6), then the Σ_1-theory of the structure $(L, \sqsubseteq, (w)_{w \in L})$ is decidable (Theorem 6.2). Note that, by [6, Theorem 3.3], this theory is undecidable if $L = \Sigma^*$.

Our first result is shown by a first-order interpretation of the structure in $(\mathbb{N}, +)$. Since $L \subseteq w_1^* \cdots w_n^*$, instead of words, one can argue about vectors $(x_1, \ldots, x_n) \in \mathbb{N}^n$ for which $w_1^{x_1} \cdots w_n^{x_n} \in L$. For the interpretation, we use the fact that semilinearity of context-free languages yields a Presburger formula expressing $w_1^{x_1} \cdots w_n^{x_n} \in L$ for $(x_1, \ldots, x_n) \in \mathbb{N}^n$. Moreover, Presburger definability of $w_1^{x_1} \cdots w_n^{x_n} \sqsubseteq w_1^{y_1} \cdots w_n^{y_n}$ for $(x_1, \ldots, x_n) \in \mathbb{N}^n$ and $(y_1, \ldots, y_n) \in \mathbb{N}^n$ is a

simple consequence of the subword relation being rational, which was observed in [12]. The first-order interpretation of our structure in $(\mathbb{N}, +)$ then enables us to employ decidability of the C+MOD-theory of the latter structure [1, 20, 5]. (Note that this decidability does not follow directly from Presburger's result since in first-order logic, one cannot make statements like "the number of witnesses $x \in \mathbb{N}$ satisfying ... is even"). A similar interpretation in $(\mathbb{N}, +)$ was used in [6] for various algorithms concerning $(\Sigma^*, \sqsubseteq, (w)_{w \in \Sigma^*})$ for fragments of FO related to bounded languages.

Our second result extends an approach from [12] for decidability of the FO^2-theory of the structure $(\Sigma^*, \sqsubseteq, (L)_{L \text{ regular}})$. The authors of [12] provide a quantifier elimination procedure showing that every unary relation FO^2-definable in this structure is regular. Our extended quantifier-elimination procedure uses the same invariant, now relying on the following two properties:

- The class of regular languages is closed under *counting* images under *unambiguous* rational relations.
  This can be shown either directly or (as we do here) using weighted automata [19].
- The proper subword, the cover, and the incomparability relation are *unambiguous* rational.

Our third result extends the decidability of the Σ_1-theory of $(\Sigma^*, \sqsubseteq)$ from [16]. In [16], decidability is a consequence of the fact that every finite partial order can be embedded into $(\Sigma^*, \sqsubseteq)$ if $|\Sigma| \geq 2$. This certainly fails for general regular languages: $(a^*, \sqsubseteq)$ can only accomodate linear orders. However, we can distinguish two cases: If L is a bounded language, then decidability of the Σ_1-theory of $(L, \sqsubseteq)$ follows from our first result. If L is not bounded, then we show that again every finite partial order embeds into $(L, \sqsubseteq)$. To this end, we first extend a well-known property of unbounded regular languages, namely that there are $x, u, v, y \in \Sigma^*$ with $x\{u, v\}^* y \subseteq L$ such that $|u| = |v|$ and $u \neq v$. We show that here, u, v can be chosen so that uv is a primitive word. We then observe that for large enough n, any embedding of the word $(uv)^{n-1}$ into $(uv)^n$ must hit either the left-most position or the right-most position in $(uv)^n$. This enables us to argue that for large enough n, sending a tuple $(t_1, \ldots, t_m) \in \{0, 1\}^m$ to $xv^{t_1}(uv)^n \cdots v^{t_m}(uv)^n y$ is in fact an embedding of $(\{0, 1\}^m, \leq)$ into $(L, \sqsubseteq)$, where $\leq$ denotes coordinate-wise comparison. Since any partial order with $\leq m$ elements embeds into $(\{0, 1\}^m, \leq)$, this completes the proof.

Regarding our fourth result, we know from [6] that decidability of the Σ_1-theory of $(L, \sqsubseteq, (w)_{w \in L})$ does not hold for every regular L: Undecidability holds already for $L = \{a, b\}^*$. Therefore, we require that every letter is frequent in L, meaning that in some automaton for L, every letter occurs in every cycle. In case L is bounded, we can again invoke our first result. If L is not bounded, we deduce from the frequency condition that for every $w \in \Sigma^*$, there are only finitely many words in L that do not have w as a subword. Removing those finitely many words preserves unboundedness, so that every finite partial order embeds in L above w. We then proceed to show that for such languages, any Σ_1-sentence is effectively equivalent to a sentence where constants are only used

to express that all variables take values above a certain word w. Since every finite partial order embeds above w, this implies decidability.

The full version of this work is available as [17].

2 Preliminaries

Throughout this paper, let Σ be some finite alphabet. A word $u = a_1 a_2 \dots a_m$ with $a_1, a_2, \dots, a_m \in \Sigma$ is a *subword* of a word $v \in \Sigma^*$ if there are words $v_0, v_1, \dots, v_m \in \Sigma^*$ with $v = v_0 a_1 v_1 a_2 v_2 \cdots a_m v_m$. In that case, we write $u \sqsubseteq v$; if, in addition, $u \neq v$, then we write $u \sqsubset v$ and call u a *proper* subword of v. If $u, w \in \Sigma^*$ such that $u \sqsubset w$ and there is no word v with $u \sqsubset v \sqsubset w$, then we say that w is a *cover* of u and write $u \sqsubset\!\!\cdot\, w$. This is equivalent to saying $u \sqsubseteq w$ and $|u| + 1 = |w|$ where $|u|$ is the length of the word u. If neither u is a subword of v nor *vice versa*, then the words u and v are *incomparable* and we write $u \parallel v$. For instance, $aa \sqsubset babbba$, $aa \sqsubset\!\!\cdot\, aba$, and $aba \parallel aabb$.

Let $\mathcal{S} = (L, (R_i)_{i \in I}, (w_j)_{j \in J})$ be a *structure*, i.e., L is a set, $R_i \subseteq L^{n_i}$ is a relation of arity n_i (for all $i \in I$), and $w_j \in L$ for all $j \in J$. Then, formulas φ of the logic C+MOD are defined by the following grammar:

$$\varphi ::= (s = t) \mid R_i(s_1, \dots, s_{n_i}) \mid \neg\varphi \mid \varphi \vee \varphi \mid \exists x\, \varphi \mid \exists^{\geq k} x\, \varphi \mid \exists^{p \bmod q} x\, \varphi$$

where $s, t, s_1, \dots, s_{n_i}$ are variables or constants w_j with $j \in J$, $i \in I$, $k \in \mathbb{N}$, and $p, q \in \mathbb{N}$ with $p < q$. We call $\exists^{\geq k}$ a *threshold counting quantifier* and $\exists^{p \bmod q}$ a *modulo counting quantifier*. The semantics of these quantifiers is defined as follows:

- $\mathcal{S} \models \exists^{\geq k} x\, \alpha$ iff $|\{w \in L \mid \mathcal{S} \models \alpha(w)\}| \geq k$
- $\mathcal{S} \models \exists^{p \bmod q} x\, \alpha$ iff $|\{w \in L \mid \mathcal{S} \models \alpha(w)\}| \in p + q\mathbb{N}$

For instance, $\exists^{0 \bmod 2} x\, \alpha$ expresses that the number of elements of the structure satisfying α is even. Then $\big(\exists^{0 \bmod 2} x\, \alpha\big) \vee \big(\exists^{1 \bmod 2} x\, \alpha\big)$ holds iff only finitely many elements of the structure satisfy α. The fragment FO+MOD of C+MOD comprises all formulas not containing any threshold counting quantifier. First-order logic FO is the set of formulas from C+MOD not mentioning any counting quantifier. Let Σ_1 denote the set of first-order formulas of the form $\exists x_1\, \exists x_2 \dots \exists x_n \colon \psi$ where ψ is quantifier-free; these formulas are also called *existential*.

The threshold quantifier $\exists^{\geq k}$ can be expressed using the existential quantifier, only. Consequently, the logics FO+MOD and C+MOD are equally expressive. The situation changes when we restrict the number of variables that can be used in a formula: Let FO+MOD^2 and C+MOD^2 denote the set of formulas from FO+MOD and C+MOD, respectively, that use the variables x and y, only. Then, the existence of ≥ 3 elements in the structure is expressible in C+MOD^2, but not in FO+MOD^2.

In this paper, we will consider the following structures:

– The largest one is $(L, \sqsubseteq, \sqsubset, (K \cap L)_{K \text{ regular}}, (w)_{w \in L})$ for some $L \subseteq \Sigma^*$. The universe of this structure is the language L, we have two binary predicates ($\sqsubseteq$ and $\sqsubset$), a unary predicate $K \cap L$ for every regular language K, and we can use every word from L as a constant.
– The other extreme is the structure $(L, \sqsubseteq)$ for some $L \subseteq \Sigma^*$ where we consider only the binary predicate $\sqsubseteq$.
– Finally, we will also prove results on the intermediate structure $(L, \sqsubseteq, (w)_{w \in L})$ that has a binary relation and any word from the language as a constant.

For any structure $\mathcal{S}$ and any of the logics $\mathcal{L}$, the *$\mathcal{L}$-theory* of $\mathcal{S}$ is the set of sentences from $\mathcal{L}$ that hold in $\mathcal{S}$.

A non-deterministic finite automaton is called *non-degenerate* if every state lies on a path from an initial to a final state. A language $L \subseteq \Sigma^*$ is *bounded* if there are a number $n \in \mathbb{N}$ and words $w_1, w_2, \ldots, w_n \in \Sigma^*$ such that $L \subseteq w_1^* w_2^* \cdots w_n^*$. Otherwise, it is *unbounded*.

For a monoid M, a subset $S \subseteq M$ is called *rational* if it is a homomorphic image of a regular language. In other words, there exists an alphabet Δ, a regular $R \subseteq \Delta^*$, and a homomorphism $h \colon \Delta^* \to M$ with $S = h(R)$. In particular, if Σ_1, Σ_2 are alphabets and $M = \Sigma_1^* \times \Sigma_2^*$, then a subset $S \subseteq \Sigma_1^* \times \Sigma_2^*$ is rational iff there is an alphabet Δ, a regular $R \subseteq \Delta^*$, and homomorphisms $h_i \colon \Delta^* \to \Sigma_i^*$ with $S = \{(h_1(w), h_2(w)) \mid w \in R\}$. This fact is known as *Nivat's theorem* [2].

For an alphabet Γ, a word $w \in \Gamma^*$, and a letter $a \in \Gamma$, let $|w|_a$ denote the number of occurrences of the letter a in the word w. The *Parikh vector* of w is the tuple $\Psi_\Gamma(w) = (|w|_a)_{a \in \Gamma} \in \mathbb{N}^\Gamma$. Note that Ψ_Γ is a homomorphism from the free monoid Γ^* onto the additive monoid $(\mathbb{N}^\Gamma, +)$.

3 The FO+MOD-theory with regular predicates

The aim of this section is to prove that the full FO+MOD-theory of the structure

$$(L, \sqsubseteq, \sqsubset, (K \cap L)_{K \text{ regular}}, (w)_{w \in L})$$

is decidable for L bounded and context-free. This is achieved by interpreting this structure in $(\mathbb{N}, +)$, i.e., in Presburger arithmetic whose FO+MOD-theory is known to be decidable [1, 20, 5]. We start with three preparatory lemmas.

Lemma 3.1. *Let $K \subseteq \Sigma^*$ be context-free, $w_1, \ldots, w_n \in \Sigma^*$, and $g \colon \mathbb{N}^n \to \Sigma^*$ be defined by $g(\overline{m}) = w_1^{m_1} w_2^{m_2} \cdots w_n^{m_n}$ for all $\overline{m} = (m_1, m_2, \ldots, m_n) \in \mathbb{N}^n$. The set $g^{-1}(K) = \{\overline{m} \in \mathbb{N}^n \mid g(\overline{m}) \in K\}$ is effectively semilinear.*

Proof. Let $\Gamma = \{a_1, a_2, \ldots, a_n\}$ be an alphabet and define the monoid homomorphism $f \colon \Gamma^* \to \Sigma^*$ by $f(a_i) = w_i$ for all $i \in [1, n]$.

Since the class of context-free languages is effectively closed under inverse homomorphisms and under intersections with regular languages, the language

$$L = f^{-1}(K) \cap a_1^* a_2^* \cdots a_n^* = \{u \in a_1^* a_2^* \cdots a_n^* \mid f(u) \in K\}$$

is effectively context-free. Its Parikh image $\Psi_\Gamma(L) \subseteq \mathbb{N}^n$ is effectively semilinear [18]. Moreover, $\Psi_\Gamma(L)$ equals the set $g^{-1}(K)$ from the lemma. □

Lemma 3.2. *Let $w_1, \ldots, w_n \in \Sigma^*$ and $g\colon \mathbb{N}^n \to \Sigma^*$ be defined by $g(\overline{m}) = w_1^{m_1} w_2^{m_2} \cdots w_n^{m_n}$ for all $\overline{m} = (m_1, m_2, \ldots, m_n) \in \mathbb{N}^n$. The set $\{(\overline{m}, \overline{n}) \in \mathbb{N}^n \times \mathbb{N}^n \mid g(\overline{m}) \sqsubseteq g(\overline{n})\}$ is semilinear.*

Proof. Let $\Gamma = \{a_1, a_2, \ldots, a_n\}$ be an alphabet and define the monoid homomorphism $f\colon \Gamma^* \to \Sigma^*$ by $f(a_i) = w_i$ for all $i \in [1, n]$. One first shows that

$$S_2 = \{(u, v) \mid u, v \in a_1^* a_2^* \ldots a_n^*,\ f(v) \sqsubseteq f(v)\}$$

is rational. We now employ Nivat's theorem. It tells us that there are a regular language R over some alphabet Δ and two homomorphisms $h_1, h_2\colon \Delta^* \to \Gamma^*$ so that we can write $S_2 = \{\big(h_1(w), h_2(w)\big) \mid w \in R\}$. Since R is regular, its Parikh-image $\Psi_\Delta(R) = \{\Psi_\Delta(w) \mid w \in R\}$ is semilinear [18]. There are monoid homomorphisms $p_1, p_2\colon \mathbb{N}^\Delta \to \mathbb{N}^n$ with $\Psi_\Gamma(h_i(w)) = p_i(\Psi_\Delta(w))$ for all $i \in \{1, 2\}$ and $w \in \Delta^*$. With these, the image $H = \{\big(p_1(\Psi_\Delta(w)), p_2(\Psi_\Delta(w))\big) \mid w \in R\}$ of the set $\Psi_\Delta(R)$ under the monoid homomorphism $(p_1, p_2)\colon \mathbb{N}^\Delta \to \mathbb{N}^n \times \mathbb{N}^n$ is semilinear. It turns out that this set equals the set from the lemma. □

Lemma 3.3. *Let $w_1, w_2, \ldots, w_n \in \Sigma^*$, $L \subseteq w_1^* w_2^* \cdots w_n^*$ be context-free, and $g\colon \mathbb{N}^n \to \Sigma^*$ be defined by $g(\overline{m}) = w_1^{m_1} w_2^{m_2} \cdots w_n^{m_n}$ for every tuple $\overline{m} = (m_1, m_2, \ldots, m_n) \in \mathbb{N}^n$. Then there exists a semilinear set $U \subseteq \mathbb{N}^n$ such that g maps U bijectively onto L.*

Proof. The set U contains, for each $u \in L$, the lexicographically minimal tuple $\overline{m} \in \mathbb{N}^n$ with $g(\overline{m}) = u$. Then, Lemmas 3.1 and 3.2 and the closure of the class of semilinear sets under first-order definitions imply the required properties. □

Now we can prove the main result of this section.

Theorem 3.4. *Let $L \subseteq \Sigma^*$ be context-free and bounded. Then the* FO+MOD*-theory of $(L, \sqsubseteq, \sqsubset, (K \cap L)_{K \text{ regular}}, (w)_{w \in L})$ is decidable.*

Proof. It suffices to prove the decidability for the structure $\mathcal{S} = (L, \sqsubseteq, (K \cap L)_{K \text{ regular}})$ since the theory of the structure from the theorem can be reduced to that of $\mathcal{S}$ ($x \sqsubset y$ gets replaced by its definition and $x \theta w$ by $\exists y\colon y \in \{w\} \wedge x \theta y$ where θ is any binary relation symbol).

Since L is bounded, there are words $w_1, w_2, \ldots, w_n \in \Sigma^*$ such that L is included in $w_1^* w_2^* \cdots w_n^*$. For an n-tuple $\overline{m} = (m_1, m_2, \ldots, m_n) \in \mathbb{N}^n$ we define $g(\overline{m}) = w_1^{m_1} w_2^{m_2} \cdots w_n^{m_n} \in \Sigma^*$.

1. By Lemma 3.3, there is a semilinear set $U \subseteq \mathbb{N}^n$ that is mapped by g bijectively onto L.
2. The set $\{(\overline{m}, \overline{n}) \mid g(\overline{m}) \sqsubseteq g(\overline{n})\}$ is semilinear by Lemma 3.2.
3. For any regular language $K \subseteq \Sigma^*$ the set $\{\overline{m} \in \mathbb{N}^n \mid g(\overline{m}) \in K\} \subseteq \mathbb{N}^n$ is effectively semilinear by Lemma 3.1.

From these semilinear sets, we obtain first-order formulas $\lambda(\overline{x})$, $\sigma(\overline{x}, \overline{y})$, and $\kappa_K(\overline{x})$ in the language of $(\mathbb{N}, +)$ such that, for any $\overline{m}, \overline{n} \in \mathbb{N}^n$, we have

1. $(\mathbb{N},+) \models \lambda(\overline{m}) \iff \overline{m} \in U$,
2. $(\mathbb{N},+) \models \sigma(\overline{m},\overline{n}) \iff g(\overline{m}) \sqsubseteq g(\overline{n})$, and
3. $(\mathbb{N},+) \models \kappa_K(\overline{m}) \iff g(\overline{m}) \in K$.

One then defines, from an FO+MOD-formula $\varphi(x_1,\ldots,x_k)$ in the language of $\mathcal{S}$, an FO+MOD-formula $\varphi'(\overline{x_1},\ldots,\overline{x_k})$ in the language of $(\mathbb{N},+)$ such that

$$(\mathbb{N},+) \models \varphi'(\overline{m_1},\ldots,\overline{m_k}) \iff \mathcal{S} \models \varphi(g(\overline{m_1}),\ldots,g(\overline{m_k}))\,.$$

(This construction can be found in the full version [17] and increases the formula size at least exponentially.)

Consequently, any sentence φ from FO+MOD in the language of $\mathcal{S}$ is translated into an equivalent sentence φ' in the language of $(\mathbb{N},+)$. By [1, 20, 5], validity of the sentence φ' in $(\mathbb{N},+)$ is decidable. □

4 The C+MOD^2-theory with regular predicates

It is the aim of this section to show that the C+MOD^2-theory of the structure $(L, \sqsubseteq, \sqsubset, (K \cap L)_{K \text{ regular}}, (w)_{w\in L})$ is decidable for any regular language L. To this aim, we first show that the C+MOD^2-theory of

$$\mathcal{S} = (\Sigma^*, \sqsubseteq, \sqsubset, (L)_{L \text{ regular}})$$

is decidable. This decidability proof extends the proof from [12] for the decidability of the FO^2-theory of $(\Sigma^*, \sqsubseteq, (L)_{L \text{ regular}})$. It provides a quantifier-elimination procedure (see Section 4.3) that relies on the following two properties:

1. The class of regular languages is closed under *counting* images under *unambiguous* rational relations (Section 4.2) and
2. the proper subword, the cover, and the incomparability relation are *unambiguous* rational (Section 4.1).

4.1 Unambiguous rational relations

Recall that, by Nivat's theorem, a relation $R \subseteq \Sigma^* \times \Sigma^*$ is rational if there exist an alphabet Γ, a homomorphism $h\colon \Gamma^* \to \Sigma^* \times \Sigma^*$, and a regular language $S \subseteq \Gamma^*$ such that h maps S surjectively onto R. We call R an *unambiguous rational relation* if, in addition, h maps S *injectively* (and therefore bijectively) onto R. Note that these are precisely the relations accepted by unambiguous 2-tape-automata.

While the class of rational relations is closed under unions, this is not the case for unambiguous rational relations (e.g., $R = \{(a^m b a^n, a^m) \mid m,n \in \mathbb{N}\} \cup \{(a^m b a^n, a^n) \mid m,n \in \mathbb{N}\}$ is the union of unambiguous rational relations but not unambiguous). But it is closed under *disjoint* unions.

Lemma 4.1. *For any alphabet Σ, the cover relation $\sqsubset$ and the relation $\sqsubset \setminus \sqsubset$ are unambiguous rational.*

Proof. For $i \in \{1,2\}$, let $\Sigma_i = \Sigma \times \{i\}$ and $\Gamma = \Sigma_1 \cup \Sigma_2$. Furthermore, let the homomorphism $\mathrm{proj}_i \colon \Gamma^* \to \Sigma^*$ be defined by $\mathrm{proj}_i(a,i) = a$ and $\mathrm{proj}_i(a, 3{-}i) = \varepsilon$ for all $a \in \Sigma$. Finally, let the homomorphism $\mathrm{proj} \colon \Gamma^* \to \Sigma^* \times \Sigma^*$ be defined by $\mathrm{proj}(w) = (\mathrm{proj}_1(w), \mathrm{proj}_2(w))$.

- The regular language

$$\mathrm{Sub} = \left(\bigcup_{a \in \Sigma} \Big(\big(\Sigma_2 \setminus \{(a,2)\}\big)^* (a,2)\,(a,1) \Big) \right)^* {\Sigma_2}^* .$$

  is mapped bijectively onto the subword relation.
- Let S be the regular language of words from Sub with precisely one more occurrence of letters from Σ_2 than from Σ_1. Then S is mapped bijectively onto the relation $\sqsubset\!\cdot$, hence this relation is unambiguous rational.
- Similarly, let S' denote the regular language of all words from Sub with at least two more occurrences of letters from Σ_2 than from Σ_1. It is mapped bijectively onto the relation $\sqsubset \setminus \sqsubset\!\cdot$, i.e., $\sqsubset \setminus \sqsubset\!\cdot$ is unambiguous rational. ◻

Lemma 4.2. *For any alphabet Σ, the incomparability relation*

$$\| \;=\; \{(u,v) \in \Sigma^* \times \Sigma^* \mid \textit{neither } u \sqsubseteq v \textit{ nor } v \sqsubseteq u\}$$

is unambiguous rational.

Proof. We will show that the following three relations are unambiguous rational:

1. $R_1 = \{(u,v) \mid |u| < |v| \text{ and not } u \sqsubseteq v\}$,
2. $R_2 = \{(u,v) \mid |u| = |v| \text{ and } u \neq v\}$, and
3. $R_3 = \{(u,v) \mid |u| > |v| \text{ and not } v \sqsubseteq u\}$.

The result follows since $\|$ is the disjoint union of these relations. Let Σ_i, Γ, proj_i, and proj be defined as in the previous proof. First, the regular language

$$\mathrm{Inc}_2 = (\Sigma_2\Sigma_1)^* \cdot \{(a,2)(b,1) \mid a,b \in \Sigma, a \neq b\} \cdot (\Sigma_2\Sigma_1)^* .$$

is mapped by proj bijectively onto R_2.

From [12, Lemma 5.2], we learn that $(u,v) \in R_1 \cup R_2$ if, and only if,

- $u = a_1 a_2 \ldots a_\ell u'$ for some $\ell \geq 1$, $a_1, \ldots, a_\ell \in \Sigma$, $u' \in \Sigma^*$, and
- $v \in (\Sigma \setminus \{a_1\})^* a_1 \, (\Sigma \setminus \{a_2\})^* a_2 \cdots (\Sigma \setminus \{a_{\ell-1}\})^* a_{\ell-1} \, (\Sigma \setminus \{a_\ell\})^+ v'$ for some word $v' \in \Sigma^*$ with $|u'| = |v'|$.

Consequently, proj maps the following language bijectively onto $R_1 \cup R_2$:

$$\mathrm{Inc}_{1,2} = \left(\bigcup_{a \in \Sigma} \Big(\big(\Sigma_2 \setminus \{(a,2)\}\big)^* (a,2)(a,1) \Big) \right)^* \cdot \bigcup_{a \in \Sigma} \Big(\big(\Sigma_2 \setminus \{(a,2)\}\big)^+ (a,1) \Big) \cdot (\Sigma_2\Sigma_1)^*$$

and since $\mathrm{Inc}_2 \subseteq \mathrm{Inc}_{1,2}$, proj maps $\mathrm{Inc}_1 = \mathrm{Inc}_{1,2} \setminus \mathrm{Inc}_2$ bijectively onto R_1. The claim regarding R_3 follows analogously. ◻

4.2 Closure properties of the class of regular languages

Let $R \subseteq \Sigma^* \times \Sigma^*$ be an unambiguous rational relation and $L \subseteq \Sigma^*$ a regular language. We want to show that the languages of all words $u \in \Sigma^*$

$$\text{with } |\{v \in L \mid (u,v) \in R\}| \geq k \tag{1}$$

$$(\text{with } |\{v \in L \mid (u,v) \in R\}| \in p + q\mathbb{N}, \text{ respectively}) \tag{2}$$

are effectively regular for all $k \in \mathbb{N}$ and all $0 \leq p < q$, respectively (this does not hold for arbitrary rational relations). It is straightforward to work out direct automata constructions for this. However, the full details of this are somewhat cumbersome. Instead, we provide a proof via weighted automata, which enables us to split the two constructions into several simple steps.

Let S be a semiring. A function $r\colon \Sigma^* \to S$ is *realizable over* S if there are $n \in \mathbb{N}$, $\lambda \in S^{1\times n}$, a homomorphism $\mu\colon \Sigma^* \to S^{n\times n}$, and $\nu \in S^{n\times 1}$ with $r(w) = \lambda \cdot \mu(w) \cdot \nu$ for all $w \in \Sigma^*$. The triple (λ, μ, ν) is a *presentation of dimension* n or a *weighted automaton for* r.

In the following, we consider the semiring $\mathbb{N}^\infty$, i.e., the set $\mathbb{N} \cup \{\infty\}$ together with the commutative operations $+$ and $\cdot$ (with $x + \infty = \infty$ for all $x \in \mathbb{N} \cup \{\infty\}$, $x \cdot \infty = \infty$ for all $x \in (\mathbb{N} \cup \{\infty\}) \setminus \{0\}$, and $0 \cdot \infty = 0$). Sometimes, we will argue about sums of infinitely many elements from $\mathbb{N}^\infty$, which are defined as expected.

Proposition 4.3. *Let Γ and Σ be alphabets, $f\colon \Gamma^* \to \Sigma^*$ a homomorphism, and $\chi\colon \Gamma^* \to \mathbb{N}^\infty$ a realizable function over $\mathbb{N}^\infty$. Then the following function r is effectively realizable over $\mathbb{N}^\infty$:*

$$r = \chi \circ f^{-1}\colon \Sigma^* \to \mathbb{N}^\infty\colon u \mapsto \sum_{\substack{w \in \Gamma^* \\ f(w) = u}} \chi(w)$$

Proof. The homomorphism f can be written as $f = f_2 \circ f_1$ where $f_1\colon \Gamma^* \to \Gamma^*$ is non-expanding (i.e., $f_1(a) \in \Gamma \cup \{\varepsilon\}$ for all $a \in \Gamma$) and $f_2\colon \Gamma^* \to \Sigma^*$ is non-erasing (i.e., $f_2(a) \in \Sigma^+$ for all $a \in \Gamma$). Then $r = (\chi \circ f_1^{-1}) \circ f_2^{-1}$. Then $\chi' = \chi \circ f_1^{-1}$ is effectively realizable by [3, Lemma 2.2(b)].

Let (λ, μ, ν) be a presentation of dimension n for χ'. For $\sigma \in \Sigma \cup \{\varepsilon\}$, set $\Gamma_\sigma = \{b \in \Gamma \mid f_2(b) = \sigma\}$. Furthermore, define the matrix $M \in (\mathbb{N}^\infty)^{n\times n}$ by

$$M_{ij} = \begin{cases} \infty & \text{if there is } w \in \Gamma_\varepsilon^* \text{ with } n < |w| \leq 2n \text{ and } \mu(w)_{ij} > 0 \\ \sum_{w \in \Gamma_\varepsilon^{\leq n}} \mu(w)_{ij} & \text{otherwise.} \end{cases}$$

Then $M_{ij} = \sum_{w \in \Gamma_\varepsilon^*} \mu(w)_{ij}$ for all $i, j \in [1, n]$. Setting $\lambda' = \lambda \cdot M$ and

$$\mu'(a) = \sum_{b \in \Gamma_a} \big(\mu(b) \cdot M\big) \text{ for all } a \in \Sigma$$

defines the presentation (λ', μ', ν) for the function $r = \chi' \circ f_2^{-1}$. □

Lemma 4.4. *Let $R \subseteq \Sigma^* \times \Sigma^*$ be an unambiguous rational relation and $L \subseteq \Sigma^*$ be regular. Then the following function r is effectively realizable over $\mathbb{N}^\infty$:*

$$r\colon \Sigma^* \to \mathbb{N}^\infty\colon u \mapsto |\{v \in L \mid (u,v) \in R\}|$$

Proof. Since R is unambiguous rational, so is $R \cap (\Sigma^* \times L)$, i.e., there are an alphabet Γ, homomorphisms $f, g\colon \Gamma^* \to \Sigma^*$, and a regular language $S_L \subseteq \Gamma^*$ such that

$$(f,g)\colon \Gamma^* \to \Sigma^* \times \Sigma^*\colon w \mapsto \big(f(w), g(w)\big)$$

maps S_L bijectively onto $R \cap (\Sigma^* \times L)$. Since S_L is regular, its characteristic function χ is effectively realizable by [19, Prop. 3.12]. One then shows that r is the function $\chi \circ f^{-1}$ as in Proposition 4.3. □

We now come to the main result of this section.

Proposition 4.5. *Let $R \subseteq \Sigma^* \times \Sigma^*$ be an unambiguous rational relation and $L \subseteq \Sigma^*$ be regular. Then, for $k \in \mathbb{N}$ and for $p, q \in \mathbb{N}$ with $p < q$, the set H of words w satisfying* (1) *and* (2)*, respectively, is effectively regular.*

Let R denote the rational relation mentioned before Lemma 4.1. Then a word $a^m b a^n$ has ≥ 2 "R-partners" iff it has an even number of "R-partners" iff $m \neq n$. Hence, the above proposition does not hold for arbitrary rational relations.

Proof. Let r be the function from Lemma 4.4. Setting $x \equiv y$ iff $x = y$ or $k \leq x, y < \infty$ defines a congruence $\equiv$ on $\mathbb{N}^\infty$. Then $S_k^\infty = \mathbb{N}^\infty/\equiv$ is a finite semiring and the function $s\colon \Sigma^* \to S_k^\infty\colon u \mapsto [r(u)]$ is effectively realizable. Since the semiring S_k^∞ is finite, the "level sets" $s^{-1}([i]) = \{u \in \Sigma^* \mid s(u) \equiv i\}$ are effectively regular by [19, Prop. 4.5]. Since $s^{-1}([k]) \cup s^{-1}([\infty])$ is the language of words u satisfying (1), the first result follows.

For the second language, we consider the congruence $\equiv\ \subseteq \mathbb{N}^\infty \times \mathbb{N}^\infty$ with $x \equiv y$ iff $x = y$ or $q \leq x, y < \infty$ and $x - y \in q\mathbb{N}$. □

4.3 Quantifier elimination for C+MOD^2

Our decision algorithm employs a quantifier alternation procedure, i.e., we will transform an arbitrary formula into an equivalent one that is quantifer-free. As usual, the heart of this procedure handles formulas $\psi = \mathrm{Q} y\, \varphi$ where Q is a quantifier and φ is quantifier-free. Since the logic C+MOD^2 has only two variables, any such formula ψ has at most one free variable. In other words, it defines a language K. The following lemma shows that this language is effectively regular, such that ψ is equivalent to the quantifier-free formula $x \in K$.

Lemma 4.6. *Let $\varphi(x,y)$ be a quantifier-free formula from* C+MOD^2 *in the language of the structure $\mathcal{S} = (\Sigma^*, \sqsubseteq, \sqsubset, (L)_{L \text{ regular}})$. Then the sets*

$$\{x \in \Sigma^* \mid \mathcal{S} \models \exists^{\geq k} y\, \varphi\} \text{ and } \{x \in \Sigma^* \mid \mathcal{S} \models \exists^{p \bmod q} y\, \varphi\}$$

are effectively regular for all $k \in \mathbb{N}$ and all $p, q \in \mathbb{N}$ with $p < q$.

Proof. Since φ is quantifier-free, we can rewrite it into a Boolean combination of formulas of the form $x \in K$ and $y \in L$ for some regular languages K and L, $x \sqsubseteq y$ and $y \sqsubseteq x$, and $x \mathbin{\sqsubset\!\!\cdot} y$ and $y \mathbin{\sqsubset\!\!\cdot} x$.

There are six possible relations between the two variables x and y in the partial order: we can have $x = y$, $x \mathbin{\sqsubset\!\!\cdot} y$ or *vice versa*, $x \sqsubset y \wedge \neg x \mathbin{\sqsubset\!\!\cdot} y$ or *vice versa*, or $x \parallel y$. Let $\theta_i(x, y)$ for $1 \leq i \leq 6$ be formulas describing these relations.

Hence φ is equivalent to $\bigvee_{1 \leq i \leq 6} (\theta_i \wedge \varphi)$. In this formula, any occurrence of φ appears in conjunction with precisely one of the formulas θ_i. Depending on this formula θ_i (i.e., the relation between x and y), we can simplify φ to φ_i by replacing the atomic subformulas that compare x and y by true or false. As a result, the formula φ is equivalent to $\bigvee_{1 \leq i \leq 6} (\theta_i \wedge \varphi_i)$ where the formulas φ_i are Boolean combinations of formulas of the form $x \in K$ and $y \in L$ for some regular languages K and L.

Now let $k \in \mathbb{N}$. Since the formulas θ_i are mutually exclusive, we get

$$\exists^{\geq k} y \, \varphi \equiv \exists^{\geq k} y \bigvee_{1 \leq i \leq 6} (\theta_i \wedge \varphi_i) \equiv \bigvee_{(*)} \bigwedge_{1 \leq i \leq 6} \exists^{\geq k_i} y \, (\theta_i \wedge \varphi_i)$$

where the disjunction $(*)$ extends over all $(k_1, \ldots, k_6) \in \mathbb{N}^6$ with $\sum_{1 \leq i \leq 6} k_i = k$.

Hence it suffices to show that

$$\{x \in \Sigma^* \mid \exists^{\geq k} y \, (\theta_i \wedge \varphi)\} \tag{3}$$

is effectively regular for all $1 \leq i \leq 6$, all $k \in \mathbb{N}$, and all Boolean combinations φ of formulas of the form $x \in K$ and $y \in L$ where K and L are regular languages. We can find regular languages K_M and L_M and a finite set I such that φ is equivalent to $\bigvee_{M \in I} (x \in K_M \wedge y \in L_M)$ and such that this disjunction is exclusive. Hence the set from (3) equals the union of the sets

$$\{x \in \Sigma^* \mid \exists^{\geq k} y \, (\theta_i \wedge x \in K_M \wedge y \in L_M)\} = K_M \cap \underbrace{\{x \in \Sigma^* \mid \exists^{\geq k} y \in L_M \colon \theta_i\}}_{H_M}$$

for $M \in I$. The set H_M is effectively regular by Proposition 4.5 and Lemmas 4.1 and 4.2. Since the language in the claim of the lemma is a Boolean combination of such sets, the first claim is demonstrated; the second follows similarly. □

The only atomic formulas with a single variable x are $x \in L$ with L regular, $x = x$, $x \sqsubseteq x$ (which are equivalent to $x \in \Sigma^*$), and $x \mathbin{\sqsubset\!\!\cdot} x$ (which is equivalent to $x \in \emptyset$). Hence, any quantifier-free formula with a single free variable x is a Boolean combination of statements of the form $x \in L$. Lemma 4.6 thus implies:

Theorem 4.7. *Let $\mathcal{S} = (\Sigma^*, \sqsubseteq, \mathbin{\sqsubset\!\!\cdot}, (L)_{L \text{ regular}})$. Let $\varphi(x)$ be a formula from* C+MOD^2*. Then the set $\{x \in \Sigma^* \mid \mathcal{S} \models \varphi\}$ is effectively regular.*

Corollary 4.8. *Let $L \subseteq \Sigma^*$ be a regular language. Then the* C+MOD^2*-theory of the structure $\mathcal{S}_L = (L, \sqsubseteq, \mathbin{\sqsubset\!\!\cdot}, (K \cap L)_{K \text{ regular}}, (w)_{w \in L})$ is decidable.*

Proof. Let $\varphi \in \mathrm{C{+}MOD}^2$ be a sentence. We build φ_L by (1) restricting all quantifications to L, (2) replace $x\theta w$ by $\exists y\colon y \in \{w\} \wedge x\theta y$, and dually for $y\theta w$ for all $w \in L$ and all binary relations θ.

With $\mathcal{S}$ the structure from Theorem 4.7, we obtain $\mathcal{S} \models \varphi_L \iff \mathcal{S}_L \models \varphi$. By Theorem 4.7, the language $\{x \mid \mathcal{S} \models \varphi_L\}$ is regular (since φ_L is a sentence, it is $\emptyset$ or Σ^*). Hence φ_L holds iff this set is nonempty, which is decidable. □

5 The Σ_1-theory

In this section, we study for which regular languages L the Σ_1-theory of the structure $(L, \sqsubseteq)$ is decidable. If L is bounded, then decidability follows from Theorem 3.4. In the case of $(\Sigma^*, \sqsubseteq)$, decidability is known as well [16]. Here, we prove decidability for every regular language L. Note that in terms of quantifier block alternation, this is optimal: The Σ_2-theory is undecidable already in the simple case of $(\{a, b\}^*, \sqsubseteq)$ [6].

Theorem 5.1. *For every regular $L \subseteq \Sigma^*$, the Σ_1-theory of $(L, \sqsubseteq)$ is decidable.*

Observe that very generally, the Σ_1-theory of a partially ordered set $(P, \leq)$ is decidable if every finite partial order embeds into $(P, \leq)$: In that case, a formula with n variables is satisfied in $(P, \leq)$ if and only if it is satisfied for some finite partial order with at most n elements. This is used to obtain decidability for the case $L = \Sigma^*$ with $|\Sigma| \geq 2$ in [16].

As mentioned above, if L is bounded, decidability follows from Theorem 3.4. If L is unbounded, it is well-known that there is a subset $x\{p, q\}^*y \subseteq L$ such that $|p| = |q|$ and $p \neq q$ (see Lemma 5.2). Since in that case, the monoids $(\{a, b\}^*, \cdot)$ and $(\{p, q\}^*, \cdot)$ are isomorphic, it is tempting to assume that $(\{a, b\}^*, \sqsubseteq)$ embeds into $(\{p, q\}^*, \sqsubseteq)$ and thus into $(x\{p, q\}^*y, \sqsubseteq)$. However, that is not the case. If $L = \{ab, ba\}^*$, then the downward closure of any infinite subset of L includes all of L. Since, on the other hand, $(\{a, b\}^*, \sqsubseteq)$ has infinite downward closed strict subsets such as a^*, it cannot embed into $(L, \sqsubseteq)$. Nevertheless, the rest of this section demonstrates that every finite partial order embeds into $(L, \sqsubseteq)$ whenever L is an unbounded regular language. By the previous paragraph, this implies Theorem 5.1.

We recall a well-known property of unbounded regular languages.

Lemma 5.2. *If $L \subseteq \Sigma^*$ is not bounded, then there are $x, y, p, q \in \Sigma^*$ such that $|p| = |q|$, $p \neq q$, and $x\{p, q\}^*y \subseteq L$.*

Proof. Let A be any non-degenerate deterministic finite automaton accepting L. Then at least one strongly connected component of A is not a cycle since otherwise, L would be bounded. Hence, there is a state s and prefix-incomparable words u, v, each of which is read on a cycle starting in s. Since u and v are prefix-incomparable, the words $p = uv$ and $q = vu$ are distinct, but equally long. Since A is non-degenerate, there are words $x, y \in \Sigma^*$ with $x\{p, q\}^*y \subseteq L$. □

To have some control over how words can embed, we prove a stronger version of Lemma 5.2. Two words $p, q \in \Sigma^*$ are *conjugate* if there are $x, y \in \Sigma^*$ with $p = xy$ and $q = yx$. A word $p \in \Sigma^*$ is *primitive* if there is no $q \in \Sigma^*$ with $p \in qq^+$.

Proposition 5.3. *For every unbounded regular language $L \subseteq \Sigma^*$, there are $x, u, v, y \in \Sigma^*$ such that $|u| = |v|$, the word uv is primitive, and $x\{u, v\}^* y \subseteq L$.*

Proof. Since L is unbounded and regular, Lemma 5.2 yields words $x, y, p, q \in \Sigma^*$ with $|p| = |q|$, $p \neq q$, and $x\{p, q\}^* y \subseteq L$. Then the words $r = pq$ and $s = pp$ are not conjugate, because every conjugate of a square is a square. Moreover, $|r| = |s|$, and $x\{r, s\}^* y \subseteq x\{p, q\}^* y \subseteq L$. Let $n = |r|$, $u = rs^{n-1}$, and $v = s^n$. Towards a contradiction, suppose $uv = rs^{2n-1}$ is not primitive. Then there is a word $w \in \Sigma^*$ with $rs^{2n-1} \in ww^+$. Depending on whether $|w| \geq n$ or $|w| < n$, we have $n \leq |w^t| \leq n^2$ either for $t = 1$ or for $t = n$. It follows that r is a prefix of w^t and that w^t is a suffix of s^n, implying that r is a factor of s^n. Since r and s are not conjugate, this is impossible. □

We are now ready to describe how to embed a finite partial order into $(L, \sqsubseteq)$. Observe that every finite partial order with m elements embeds into $(\{0,1\}^m, \leq)$ where $\leq$ is the componentwise order. Hence, it suffices to embed this partial order into $(\{u, v\}^*, \sqsubseteq)$. We do this as follows. Let $n = |uv| + m + 3$ and define, for a tuple $t = (t_1, \ldots, t_m) \in \{0,1\}^m$,

$$\varphi_m(t_1, \ldots, t_m) = v^{t_1}(uv)^n \cdots v^{t_m}(uv)^n .$$

Then, clearly, $s \leq t$ implies $\varphi_m(s) \sqsubseteq \varphi_m(t)$. The converse requires a careful analysis of how prefixes of $\varphi_m(s)$ can embed into prefixes of $\varphi_m(t)$. For $x, y \in \Sigma^*$, we write $x \hookrightarrow y$ if x, but no word xa with $a \in \Sigma$ is a subword of y. In other words, $x \hookrightarrow y$ if x is a *prefix-maximal subword of y*. This gives us a criterion for non-embeddability: If x has a strict prefix x_0 with $x_0 \hookrightarrow y$, then certainly $x \not\sqsubseteq y$. In this case, the word x_1 with $x = x_0 x_1$ is called *residue*. We show the following:

Lemma 5.4. *Let $u, v \in \Sigma^*$ be words such that $|u| = |v|$ and uv is primitive. Then, for all $\ell, n \in \mathbb{N}$ with $n > |uv| + \ell + 2$, we have*

(i) $(uv)^n \hookrightarrow v(uv)^n$,
(ii) $(uv)^\ell v(uv)^{n-\ell-1} \hookrightarrow (uv)^n$, *and*
(iii) $(uv)^{1+\ell} v(uv)^{n-\ell-2} \hookrightarrow v(uv)^n$.

Here, it is crucial to observe that for a primitive word w and $n > |w|$, any embedding of w^n into w^{n+1} must either hit the left-most or the right-most position in w^{n+1}. To prove that $s \not\leq t$ implies $\varphi_m(s) \not\sqsubseteq \varphi_m(t)$, we argue about prefixes of the form $p_i = v^{s_1}(uv)^n \cdots v^{s_i}(uv)^n$ and $q_i = v^{t_1}(uv)^n \cdots v^{t_i}(uv)^n$ for $i \in [1, m]$. If $s \not\leq t$, let $i \in [1, m]$ be the index with $s_i = 1$, $t_i = 0$ and $s_j \leq t_j$ for all $j \in [1, i-1]$. Then clearly $p_{i-1} \sqsubseteq q_{i-1}$. In fact, Lemma 5.4 (i) implies that even $p_{i-1} \hookrightarrow q_{i-1}$, since $x \hookrightarrow y$ and $x' \hookrightarrow y'$ imply $xy \hookrightarrow x'y'$. Then, by Lemma 5.4 (ii), $p_i = p_{i-1} v(uv)^{n-1}(uv)$ has a residue of uv in $q_i = q_{i-1}(uv)^n$.

To conclude $\varphi_m(s) \not\sqsubseteq \varphi_m(t)$, it remains to be shown that this can never be rectified when considering prefixes p_j and q_j for $j = i+1, \ldots, m$. To this end, Lemma 5.4 (ii) and (iii) tell us that if p_j has a residue of $(uv)^\ell$ in q_j, then the word p_{j+1} has a residue of $(uv)^\ell$ or even $(uv)^{\ell+1}$ in q_{j+1}.

6 The Σ_1-theory with constants

In this section, we study for which languages L the structure $(L, \sqsubseteq, (w)_{w\in L})$ has a decidable Σ_1-theory. From Theorem 3.4, we know that this is the case whenever L is bounded. However, there are very simple languages for which decidability is lost: If $|\Sigma| \geq 2$, then the Σ_1-theory of $(\Sigma^*, \sqsubseteq, (w)_{w\in\Sigma^*})$ is undecidable [6]. Here, we present a sufficient condition for the Σ_1-theory of $(L, \sqsubseteq, (w)_{w\in\Sigma^*})$ to be decidable.

Let $L \subseteq \Sigma^*$. We say that a letter $a \in \Sigma$ is *frequent* in L if there is a real constant $\delta > 0$ so that $|w|_a \geq \delta \cdot |w|$ for all but finitely many $w \in L$. Our sufficient condition requires that all letters be frequent in L. Note that if L is regular, this is equivalent to saying that in a non-degenerate automaton for L, every cycle contains every letter. An example of such a regular language is $\{ab, ba\}^*$.

We shall prove that this condition implies decidability of the Σ_1-theory of $(L, \sqsubseteq, (w)_{w\in\Sigma^*})$. If L is bounded, decidability already follows from Theorem 3.4. In case L is unbounded, we employ our results from section 5 to show another embeddability result. For $w \in \Sigma^*$, let $w{\uparrow} = \{u \in \Sigma^* \mid w \sqsubseteq u\}$ denote the upward closure of $\{w\}$ in $(\Sigma^*, \sqsubseteq)$. We will show that if L is unbounded, then for each $w \in \Sigma^*$, the decomposition of $L = (L \setminus w{\uparrow}) \cup (L \cap w{\uparrow})$ yields two simple parts: The set $L \setminus w{\uparrow}$ is finite and the set $L \cap w{\uparrow}$ embeds every finite partial order. This simplifies the conditions under which a Σ_1-sentence is satisfied.

Lemma 6.1. *Let $L \subseteq \Sigma^*$ be an unbounded regular language where every letter is frequent. For every $w \in \Sigma^*$, the set $L \setminus w{\uparrow}$ is finite and $L \cap w{\uparrow}$ is unbounded.*

Proof. In a non-degenerate automaton A for L, every cycle must contain every letter. Therefore, if A has n states and $v \in L$ has $|v| > n \cdot |w|$, then a computation for v must contain some state more than $|w|$ times, which implies $w \sqsubseteq v$ and hence $v \notin L \setminus w{\uparrow}$. Therefore, $L \setminus w{\uparrow}$ is finite. This implies that $L \cap w{\uparrow}$ is unbounded: Otherwise $L = (L \cap w{\uparrow}) \cup (L \setminus w{\uparrow})$ would be bounded as well. □

Theorem 6.2. *Let $L \subseteq \Sigma^*$ be an unbounded regular language where every letter is frequent. Then the Σ_1-theory of $(L, \sqsubseteq, (w)_{w\in L})$ is decidable.*

Proof. For decidability, we may assume that we are given a formula φ that is a disjunction of conjunctions of literals of the following forms (where x and y are arbitrary variables and w an arbitrary word from L):

(i) $x \sqsubseteq w$ (iii) $w \sqsubseteq x$ (v) $x \sqsubseteq y$
(ii) $x \not\sqsubseteq w$ (iv) $w \not\sqsubseteq x$ (vi) $x \not\sqsubseteq y$

Step 1. We first show that literals of types (i) and (iv) can be eliminated. To this end, we observe that for each $w \in L$, both of the sets $\{u \in L \mid u \sqsubseteq w\}$, and

$\{u \in L \mid w \not\sqsubseteq u\}$ are finite (in the latter case, this follows from Lemma 6.1). Thus, every conjunction that contains a literal $x \sqsubseteq w$ or $w \not\sqsubseteq x$, constrains x to finitely many values. Therefore, we can replace this conjunction with a disjunction of conjunctions that result from replacing x by one of these values. (Here, we might obtain literals $u \sqsubseteq v$ or $u \not\sqsubseteq v$, but those can be replaced by other equivalent formulas). We repeat this until there are no more literals of the form (i) and (iv).

Step 2. We now eliminate literals of the form (ii). Note that the language $\{u \in L \mid u \not\sqsubseteq w\}$ is upward closed in $(L, \sqsubseteq)$. Since L is regular, we can compute the finite set of minimal elements of this set. Thus, $x \not\sqsubseteq w$ is equivalent to a finite disjunction of literals of the form $w' \sqsubseteq x$. The resulting formula ψ is a disjunction of conjunction of literals of the form (iii), (v), (vi).

Step 3. To check satisfiability, we may assume that ψ is a conjunction of literals of the form (iii), (v), (vi). We can write ψ as $\gamma_1 \wedge \gamma_2$, where γ_1 is a conjunction of literals of the form (iii) and γ_2 is a conjunction of literals of the form (v) and (vi). We claim that ψ is satisfiable if and only if γ_2 is satisfiable in some partial order. The "only if" direction is trivial, so suppose γ_2 is satisfied by some finite partial order $(P, \leq)$ and let $w \in \Sigma^*$ be a concatenation of all words occurring in γ_1. By Lemma 6.1, $L \cap w{\uparrow}$ is unbounded, which implies that $(P, \leq)$ can be embedded into $(L \cap w{\uparrow}, \sqsubseteq)$ (see section 5). This means, there exists a satisfying assignment where even $w \sqsubseteq x$ for every variable x. In particular, it satisfies $\psi = \gamma_1 \wedge \gamma_2$. □

Open questions

We did not consider complexity issues. In particular, from [12], we know that the FO^2-theory of the structure $(\Sigma^*, \sqsubseteq, (w)_{w \in \Sigma^*})$ can be decided in elementary time. We are currently working out the details for the extension of this result to the $\mathrm{C{+}MOD}^2$-theory of the structure $(L, \sqsubseteq, (w)_{w \in L})$ for regular languages L. We reduced the FO+MOD-theory of the full structure (for L context-free and bounded) to the FO+MOD-theory of $(\mathbb{N}, +)$, which is known to be decidable in elementary time [5]. Our reduction increases the formula exponentially due to the need of handling statements of the form "there is an even number of pairs $(x, y) \in \mathbb{N}^2$ such that ..." It should be checked whether the proof from [5] can be extended to handle such statements in FO+MOD for $(\mathbb{N}, +)$ directly.

Finally, our results raise an interesting question: For which regular languages L does the structure $(L, \sqsubseteq, (w)_{w \in L})$ have a decidable Σ_1-theory? If every letter is frequent in L, we have decidability. For example, this applies to $L = \{ab, ba\}^*$ or $L = \{ab, baa\}^* \cup bb\{abb\}^*$. If $L = \Sigma^*$ for $|\Sigma| \geq 2$, we have undecidability [6].

References

1. H. Apelt. Axiomatische Untersuchungen über einige mit der Presburgerschen Arithmetik verwandten Systeme. *Z. Math. Logik Grundlagen Math.*, 12:131–168, 1966.

2. J. Berstel. *Transductions and context-free languages.* Teubner Studienbücher, Stuttgart, 1979.
3. M. Droste and P. Gastin. Weighted automata and weighted logics. In M. Droste, W. Kuich, and H. Vogler, editors, *Handbook of Weighted Automata*, pages 176–211. Springer, 2009.
4. A. Finkel and Ph. Schnoebelen. Well-structured transition systems everywhere! *Theoretical Computer Science*, 256:63–92, 2001.
5. P. Habermehl and D. Kuske. On Presburger arithmetic extended with modulo counting quantifiers. In *FoSSaCS'15*, Lecture Notes in Comp. Science vol. 9034, pages 375–389. Springer, 2015.
6. S. Halfon, Ph. Schnoebelen, and G. Zetzsche. Decidability, complexity, and expressiveness of first-order logic over the subword ordering. In *Proc. of the Thirty-Second Annual ACM/IEEE Symposium on Logic in Computer Science (LICS 2017)*, pages 1–12. IEEE Computer Society, 2017.
7. G. Higman. Ordering by divisibility in abstract algebras. *Proc. London Math. Soc.*, 2:326–336, 1952.
8. J. Ježek and R. McKenzie. Definability in substructure orderings. I: Finite semilattices. *Algebra Univers.*, 61(1):59–75, 2009.
9. J. Ježek and R. McKenzie. Definability in substructure orderings. III: Finite distributive lattices. *Algebra Univers.*, 61(3-4):283–300, 2009.
10. J. Ježek and R. McKenzie. Definability in substructure orderings. IV: Finite lattices. *Algebra Univers.*, 61(3-4):301–312, 2009.
11. J. Ježek and R. McKenzie. Definability in substructure orderings. II: Finite ordered sets. *Order*, 27(2):115–145, 2010.
12. P. Karandikar and Ph. Schnoebelen. Decidability in the logic of subsequences and supersequences. In Prahladh Harsha and G. Ramalingam, editors, *Proceedings of the 35th Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS'15)*, volume 45 of *Leibniz International Proceedings in Informatics*, pages 84–97. Leibniz-Zentrum für Informatik, 2015.
13. O. V. Kudinov and V. L. Selivanov. Undecidability in the homomorphic quasiorder of finite labelled forests. *J. Log. Comput.*, 17(6):1135–1151, 2007.
14. O. V. Kudinov, V. L. Selivanov, and L. V. Yartseva. Definability in the subword order. In *Programs, Proofs, Processes, 6th Conference on Computability in Europe, CiE 2010, Ponta Delgada, Azores, Portugal, June 30 - July 4, 2010. Proceedings*, Lecture Notes in Computer Science vol. 6158, pages 246–255. Springer, 2010.
15. O. V. Kudinov, V. L. Selivanov, and A. V. Zhukov. Definability in the h-quasiorder of labeled forests. *Ann. Pure Appl. Logic*, 159(3):318–332, 2009.
16. D. Kuske. Theories of orders on the set of words. *Theoretical Informatics and Applications*, 40:53–74, 2006.
17. D. Kuske and G. Zetzsche. Languages ordered by the subword order. *CoRR*, abs/1901.02194, 2019.
18. R. Parikh. On context-free languages. *Journal of the ACM*, 13(4):570–581, 1966.
19. J. Sakarovitch. Rational and recognizable series. In M. Droste, W. Kuich, and H. Vogler, editors, *Handbook of Weighted Automata*, pages 105–174. Springer, 2009.
20. N. Schweikardt. Arithmetic, first-order logic, and counting quantifiers. *ACM Trans. Comput. Log.*, 6(3):634–671, 2005.
21. R.S. Thinniyam. Definability of recursive predicates in the induced subgraph order. In *Logic and Its Applications - 7th Indian Conference, ICLA 2017, Kanpur, India, January 5-7, 2017, Proceedings*, Lecture Notes in Comp. Science vol. 10119, pages 211–223. Springer, 2017.

22. R.S. Thinniyam. Defining recursive predicates in graph orders. *Logical Methods in Computer Science*, 14(3), 2018.