On Finite-Index Indexed Grammars and Their Restrictions ☆

Flavio D'Alessandro[a,1,*], Oscar H. Ibarra[b,2], Ian McQuillan[c,3]

[a] *Department of Mathematics*
Sapienza University of Rome, 00185 Rome, Italy
and
Department of Mathematics, Boğaziçi University
34342 Bebek, Istanbul, Turkey
[b] *Department of Computer Science*
University of California, Santa Barbara, CA 93106, USA
[c] *Department of Computer Science, University of Saskatchewan*
Saskatoon, SK S7N 5A9, Canada

Abstract

The family, $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$, of languages generated by linear indexed grammars has been studied in the literature. It is known that the Parikh image of every language in $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ is semi-linear. However, there are bounded semi-linear languages that are not in $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$. Here, we look at larger families of (restricted) indexed languages and study their combinatorial and decidability properties, and their relationships.

Keywords: Indexed Languages, Finite-Index, Full Trios, Semi-linearity, Bounded Languages, ET0L Languages

1. Introduction

Indexed grammars [1, 2] are a natural generalization of context-free grammars, where variables keep stacks of indices. Although they are included in the context-sensitive languages, the languages generated by indexed grammars are quite broad as they contain some non semi-linear languages. Several restrictions have been studied that have desirable computational properties. Linear indexed grammars were first created, restricting the number of variables on the right hand side to be at most one [6]. Other restrictions include another system named exactly linear indexed grammars [8] (see also [20]), which are different than the first formalism, although both are sufficiently restricted to only generate semi-linear languages. In this paper, we only examine the first formalism of linear indexed grammars.

We study indexed grammars that are restricted to be finite-index, which is a generalization of linear indexed grammars [6]. Such grammars generate languages that inherit several properties satisfied by context-free languages CFL. Grammar systems that are *k-index* are restricted so that, for every word generated by the grammar, there is some successful derivation where at most k variables (or nonterminals) appear in every sentential form of the derivation [5, 10, 16, 18]. A system is finite-index if it is k-index for some k. It has been found that that when restricting many different types of grammar systems to be finite-index, their languages coincide. This is the case for finite-index ET0L, EDT0L, context-free programmed grammars, ordered grammars, and matrix context-free grammars.

*Corresponding author
URL: dalessan@mat.uniroma1.it (Flavio D'Alessandro), ibarra@cs.ucsb.edu (Oscar H. Ibarra), mcquillan@cs.usask.ca (Ian McQuillan)

[1]Supported by EC-FP7 Marie-Curie/TÜBITAK/Co-Funded Brain Circulation Scheme Project 2236 (Flavio D'Alessandro).
[2]Supported, in part, by NSF Grant CCF-1117708 (Oscar H. Ibarra).
[3]Supported, in part, by Natural Sciences and Engineering Research Council of Canada Grant 2016-06172 (Ian McQuillan).

We introduce the family, $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$, of languages generated by finite-index indexed grammars and a sub-family, $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$, of languages generated by uncontrolled finite-index indexed grammars, where every successful derivation has to be finite-index. The grammars generating the languages of $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ have been very recently studied under the name of *breadth-bounded grammars* [21], and it was shown that this family is a semi-linear full trio. We also study a special case of the latter, called $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$ that restricts branching productions. We then show the following:

1. All families are full trios.
2. The semi-linearity property of $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ and $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$ is extended to a bigger family, showing, more generally, that, if $\mathcal{C}$ is an arbitrary full trio of semi-linear languages and $\mathcal{L}(\mathsf{NCM})$ is the family of languages accepted by one-way deterministic reversal-bounded multicounter machines, then every language in the family
$$\{L_1 \cap L_2 : L_1 \in \mathcal{C}, L_2 \in \mathcal{L}(\mathsf{NCM})\},$$
has a semi-linear Parikh image.
3. The following conditions are equivalent for a bounded language L:
   - $L \in \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$,
   - $L \in \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$,
   - L is bounded semi-linear,
   - L can be generated by a finite-index $\mathsf{ET0L}$ system,
   - L can be accepted by a DFA augmented with reversal-bounded counters,
4. Every finite-index $\mathsf{ET0L}$ language is in $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$,
5. $\mathsf{CFL} \subset \mathcal{L}(\mathsf{IND}_{\mathsf{LIN}}) \subset \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1}) \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}}) \subset \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$,
6. Containment and equality are decidable for bounded languages in $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ and $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$.

2. Preliminaries

We assume a basic background in formal languages and automata theory [4, 9, 10, 11].

Let k be a positive integer and let $\mathbb{N}^k$ be the additive free commutative monoid of k-tuples of non negative integers. If B is a subset of $\mathbb{N}^k$, $B^{\oplus}$ denotes the submonoid of $\mathbb{N}^k$ generated by B.

An *alphabet* is a finite set of symbols, and given an alphabet A, A^* is the free monoid generated by A. An element $w \in A^*$ is called a *word*, the empty word is denoted by λ, and any $L \subseteq A^*$ is a *language*. The *length* of a word $w \in A^*$ is denoted by $|w|$, and the number of a's, $a \in A$, in w is denoted by $|w|_a$, extended to subsets X of A by $|w|_X = \sum_{a \in X} |w|_a$.

Let $A = \{a_1, \ldots, a_t\}$ be an alphabet of t letters, and let $\psi : A^* \to \mathbb{N}^t$ be the corresponding *Parikh morphism* defined by $\psi(w) = (|w|_{a_1}, \ldots, |w|_{a_t})$.

A set $B \subseteq \mathbb{N}^k$ is a *linear set* if there exist vectors $\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_n$ of $\mathbb{N}^k$ such that $B = \mathbf{b}_0 + \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}^{\oplus}$. Further, B is called a *semi-linear set* if $B = \bigcup_{i=1}^{m} B_i, m \geq 1$, for linear sets $B_1, \ldots, B_m$. A language $L \subseteq A^*$ is said to be *semi-linear* if the Parikh morphism applied to L gives a semi-linear set. A language family is said to be semi-linear if all languages in the family are semi-linear. Many known families are semi-linear, such as the regular languages and context-free languages (denoted by CFL, see [4, 9, 10, 11]), and finite-index $\mathsf{ET0L}$ languages $\mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}})$, see [16, 17]).

A language L is termed *bounded* if there exist non-empty words $u_1, \ldots, u_k$, with $k \geq 1$, such that $L \subseteq u_1^* \cdots u_k^*$. Let $\varphi : \mathbb{N}^k \to u_1^* \cdots u_k^*$ be the map defined as: for every tuple $(\ell_1, \ldots, \ell_k) \in \mathbb{N}^k$,

$$\varphi(\ell_1, \ldots, \ell_k) = u_1^{\ell_1} \cdots u_k^{\ell_k}.$$

The map φ is called the *Ginsburg map*.

Definition 1. *A bounded language $L \subseteq u_1^* \cdots u_k^*$ is said to be* bounded Ginsburg semi-linear *if there exists a semi-linear set B of $\mathbb{N}^k$ such that $\varphi(B) = L$.*

In the literature, bounded Ginsburg semi-linear has also been called just bounded semi-linear, but we will use the terminology bounded Ginsburg semi-linear henceforth in this paper.

A *full trio* is a language family closed under morphism, inverse morphism, and intersection with regular languages [4].

We will also relate our results to the languages accepted by one-way nondeterministic reversal-bounded multicounter machines (denoted by $\mathcal{L}(\mathsf{NCM})$), and to one-way deterministic reversal-bounded multicounter machines (denoted by $\mathcal{L}(\mathsf{DCM})$. These are NFAs (DFAs) augmented by a set of counters that can switch between increasing and decreasing a fixed number of times [3, 12]).

3. Restrictions on Indexed Grammars

We first recall the definition of indexed grammar introduced in [1] by following [11], Section 14.3 (see also [5] for a reference book for grammars).

Definition 2. *An indexed grammar is a 5-tuple $G = (V, T, I, P, S)$, where*

- *V, T, I are finite pairwise disjoint sets: the set of variables, terminals, and indices, respectively;*
- *P is a finite set of productions of the forms*

$$\textbf{1)}\ A \to \nu, \quad \textbf{2)}\ A \to Bf, \quad \textit{or} \quad \textbf{3)}\ Af \to \nu,$$

  where $A, B \in V$, $f \in I$ and $\nu \in (V \cup T)^*$;
- *$S \in V$ is the start variable.*

Let us now define the derivation relation $\Rightarrow_G$ of G. Let ν be an arbitrary sentential form of G,

$$u_1 A_1 \alpha_1 u_2 A_2 \alpha_2 \cdots u_k A_k \alpha_k u_{k+1},$$

with $A_i \in V, \alpha_i \in I^*, u_i \in T^*$. For a sentential form $\nu' \in (VI^* \cup T)^*$, we set $\nu \Rightarrow_G \nu'$ if one of the following three conditions holds:

1) In P, there exists a production of the form (1) $A \to w_1 C_1 \cdots w_\ell C_\ell w_{\ell+1}$, $C_j \in V, w_j \in T^*$, such that in the sentential form ν, for some i with $1 \leq i \leq k$, one has $A_i = A$ and

$$\nu' = u_1 A_1 \alpha_1 \cdots u_i (w_1 C_1 \alpha_i \cdots w_\ell C_\ell \alpha_i w_{\ell+1}) u_{i+1} A_{i+1} \alpha_{i+1} \cdots u_k A_k \alpha_k u_{k+1}.$$

2) In P, there exists a production of the form (2) $A \to Bf$ such that in the sentential form ν, for some i with $1 \leq i \leq k$, one has $A_i = A$ and $\nu' = u_1 A_1 \alpha_1 \cdots u_i (Bf\alpha_i) u_{i+1} A_{i+1} \alpha_{i+1} \cdots u_k A_k \alpha_k u_{k+1}$.

3) In P, there exists a production of the form (3) $Af \to w_1 C_1 \cdots w_\ell C_\ell w_{\ell+1}$, $C_j \in V, w_j \in T^*$, such that in the sentential form ν, for some i with $1 \leq i \leq k$, one has $A_i = A$, $\alpha_i = f\alpha_i', \alpha_i' \in I^*$, and

$$\nu' = u_1 A_1 \alpha_1 \cdots u_i (w_1 C_1 \alpha_i' \cdots w_\ell C_\ell \alpha_i' w_{\ell+1}) u_{i+1} A_{i+1} \alpha_{i+1} \cdots u_k A_k \alpha_k u_{k+1}.$$

In this case, one says that the index f is consumed.

For every $n \in \mathbb{N}$, $\Rightarrow_G^n$ stands for the n-fold product of $\Rightarrow_G$ and $\Rightarrow_G^*$ stands for the reflexive and transitive closure of $\Rightarrow_G$. The language $L(G)$ generated by G is the set $L(G) = \{u \in T^* : S \Rightarrow_G^* u\}$.

Notation and Convention. *In the sequel we will adopt the following notation and conventions for an indexed grammar G.*

- *If no ambiguity arises, the relations* $\Rightarrow_G$, $\Rightarrow_G^n$, $n \in \mathbb{N}$, *and* $\Rightarrow_G^*$ *will be simply denoted by* $\Rightarrow$, $\Rightarrow^n$, *and* $\Rightarrow^*$, *respectively.*

- *capital letters as* $A, B, ...etc$ *(as well as its indexed variants) will denote variables of* G.

- *the small letters* e, f, *as well as* f_i, *will be used to denote indices while* α, β *and* γ, *as well as its indexed variants (as for instance* α_i*), will denote arbitrary words over* I.

- *Small letters as* $a, b, c, ...etc$ *(as well as its indexed variants) will denote letters of* T *and small letters as* $u, v, w, r..., etc$ *(as well as its indexed variants) will denote words over* T.

- ν *and* μ, *as well as* ν_i *and* μ_i, *will denote arbitrary sentential forms of* G.

- *in order to shorten the notation, according to Definition 2, if* p *is a production of* G *of the form (1) or (3), we will simply write*

$$Af \to \nu, \quad f \in I \cup \{\lambda\},$$

  where it is understood that if $f = \lambda$, *the production* p *has form (1) and if* $f \in I$, *the production* p *has form (3).*

- *If* $p \in P$ *is a production of* G, *then* $\mu \Rightarrow_p \nu$ *denotes the* 1*-step derivation of* G *defined by* p;

- *a derivation of* G *of the form* $\nu_0 \Rightarrow_{p_1} \nu_1 \Rightarrow_{p_2} \cdots \Rightarrow_{p_n} \nu_n$ *will be also shortly denoted as* $\nu_0 \Rightarrow_{p_1 \cdots p_n} \nu_n$.

The following set of definitions defines the main objects studied in this paper. Let G be an indexed grammar and let $L(G)$ be the language generated by G. The first definition is from [6].

Definition 3. *We say that* G *is* linear *if the right side component of every production of* G *has at most one variable. A language* L *is said to be* linear indexed *if there exists a linear indexed grammar* G *such that* $L = L(G)$.

Definition 4. *Given an integer* $k \geq 1$, *a derivation* $\nu_0 \Rightarrow \nu_1 \Rightarrow \cdots \Rightarrow \nu_n$ *of* $G = (V, T, I, P, S)$, *is said to be of* index-k *if* $|\nu_i|_V \leq k$, *for all* i, $0 \leq i \leq n$.

Definition 5. *Given an integer* $k \geq 1$, G *is said to be* of index-k *if, for every word* $u \in L(G)$, *there exists a derivation of* u *in* G *of index-*k.

A language L *is said to be an* indexed language of index-k *if there exists an indexed grammar* G *of index-*k *such that* $L = L(G)$. *An indexed language* L *is said to be* of finite-index *if* L *is of index-*k, *for some* k.

Definition 6. *An indexed grammar* G *is said to be* uncontrolled *index-*k *if, for every derivation* $\nu_0 \Rightarrow \cdots \Rightarrow \nu_n$ *generating* $u \in L(G), |\nu_i|_V \leq k$, *for all* i, $0 \leq i \leq n$. G *is uncontrolled finite-index if* G *is uncontrolled index-*k, *for some* k. *A language* L *is said to be an* uncontrolled finite-index indexed language *if there exists an uncontrolled finite-index grammar* G *such that* $L = L(G)$.

Remark 1. *It is worth noticing that, according to Definition 5, if* G *is a grammar of index-*k_1, *then* G *is a grammar of index-*k_2, *for every integer* $k_1 \leq k_2$.

Remark 2. *It is interesting to observe that Definition 6 corresponds, in the case of context-free grammars, to the definition of nonterminal bounded grammar (cf [10], Section 5.7). We recall that nonterminal bounded grammars are equivalent to ultralinear grammars and thus provide a characterisation of the family of languages that are accepted by Finite-Turn pushdown automata.*

Finally let us denote by

- $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ the family of finite-index indexed languages;

- $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ the family of uncontrolled finite-index indexed languages;
- $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ the family of linear indexed languages.

Uncontrolled finite-index grammars have been studied under the name of breadth-bounded indexed grammars in [21], where the following result has been proved.

Theorem 7. *$\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ is a semi-linear full trio.*

(Makoto Kanazawa has pointed out to the authors that Georg Zetzsche's result – the Parikh image of every language in $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ is semilinear – can be also obtained as corollary of a result proved in his paper [15]).

The family $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ has been introduced in [6] where results of an algebraic and combinatorial nature characterize the structure of its languages. Recall that a linear indexed grammar G is said to be *right linear indexed* if, according to Definition 2, in every production p of G of the form (1) or (3), the right hand component ν of p has the form $\nu = u$, or $\nu = uB$, where $u \in T^*, B \in V$. In [1] (see also [6]), the following theorem has been proved:

Theorem 8. *If L is an arbitrary language, L is context-free if and only if there exists a right linear indexed grammar G such that $L = L(G)$.*

From this, the following is evident.

Theorem 9. $\mathsf{CFL} \subset \mathcal{L}(\mathsf{IND}_{\mathsf{LIN}}) \subset \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}}) \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$.

Indeed Theorem 8 provides the inclusion $\mathsf{CFL} \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$. The inclusions $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}}) \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}}) \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ come immediately from the definitions of the corresponding families. In [6] (see Theorem 2.8), it is shown that for an alphabet T, and a letter $\$ \notin T$, if $M_1\$M_2$, with $M_1, M_2 \subseteq T^*$, is a linear indexed language, then M_1 or M_2 is a context-free language. Let T be an alphabet with at least two letters. Let $L_1 = \{u^2 : u \in T^*\}$, and let $L_2 = \{u^2\$v^2 : u, v \in T^*\}$. One easily sees that $L_1 \in \mathcal{L}(\mathsf{IND}_{\mathsf{LIN}}) \setminus \mathsf{CFL}$, and, since $L_2 = L_1\$L_1$, by the previous remark, $L_2 \notin \mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$. On the other hand, it is easily shown that $L_2 \in \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$. More generally, one can verify that, for every $k \geq 1$, $L_k = \{u^k\$v^k : u, v \in T^*\} \in \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$.

By applying the same argument, one has that, on the alphabet $T = \{a, b, c, \$\}$, the language $L = \{a^n b^n c^n \$ a^m b^m c^m : n, m \geq 0\}$ cannot be linear indexed.

Next, closure under union and product is addressed for the family $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$.

Lemma 10. *The family $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ is closed under union and concatenation.*

Proof. Let L_1 and L_2 be indexed languages of indices k_1 and k_2 respectively, and let G_1 and G_2 be grammars

$$G_1 = (V_1, T_1, I_1, P_1, S_1), \quad G_2 = (V_2, T_2, I_2, P_2, S_2),$$

such that $L_1 = L(G_1)$ and $L_2 = L(G_2)$. Since we may rename variables and indices without changing the language generated, we assume that $V_1 \cap V_2 = I_1 \cap I_2 = \emptyset$. Moreover let S be a new variable not in $V_1 \cup V_2$.

Construct a new grammar $G = (V, T, I, P, S)$, where $V = V_1 \cup V_2 \cup \{S\}$, $I = I_1 \cup I_2$, and P is equal to $P_1 \cup P_2$, plus the two productions $S \to S_1$ and $S \to S_2$. It is easily checked that $L_1 \cup L_2 = L(G)$ and G is of index $\max\{k_1, k_2\}$.

For concatenation, let $G' = (V, T, I, P', S)$ be the grammar obtained from G, by setting P' equal to $P_1 \cup P_2$, plus the production $S \to S_1 S_2$. It is easily checked that $L_1 L_2 = L(G')$ and G' is of index $1 + \max\{k_1, k_2\}$. □

Next, we show that $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ is a full trio. As a consequence, by using Nivat's theorem for the characterisation of rational relations of free monoids (see [4], Ch. III, Thm 4.1), we will prove the fact that they are closed under rational transductions. The proof is structured using a chain of lemmas.

Lemma 11. *$\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ is closed under morphisms.*

PROOF. Let $L \in \mathcal{L}(\mathsf{IND_{FIN}})$ and let $G = (V, T, I, P, S)$ be a k-index indexed grammar such that $L = L(G)$. Let $\varphi : T^* \to (T')^*$ be a morphism where T and T' are two alphabets. Construct a new grammar G' by replacing each production of G of the form

$$Xf \to u_1 X_1 \cdots u_\ell X_\ell u_{\ell+1},$$

where $f \in I \cup \{\lambda\},\ u_i \in T^*,\ X, X_i \in V$, by the production

$$Xf \to \varphi(u_1) X_1 \cdots \varphi(u_\ell) X_\ell \varphi(u_{\ell+1}).$$

It is easily verified that the resulting grammar G' satisfies $\varphi(L) = L(G')$ and G' is a k-index grammar. □

Lemma 12. *$\mathcal{L}(\mathsf{IND_{FIN}})$ is closed under intersection with regular languages.*

PROOF. Let $G = (V, T, I, P, S)$ be a finite-index indexed grammar and let $L = L(G)$. Let R be a regular language and let $\mathcal{A}$ be a finite automaton accepting R. Our main goal is to construct a finite index grammar G' such that $L(G') = L \cap R$. Given an arbitrary word $w \in L \cap R$, on one hand, G' should generate w, and, on the other hand, should simulate in the automaton $\mathcal{A}$ the recognition process of w. Thus, a relevant step in the construction of G', is to match the 1-step derivations of G with the atomic transitions of $\mathcal{A}$. It is thus convenient to rewrite the productions of G in a suitable canonical form. For this reason, the following claim is needed.

Claim 1. *There exists an indexed grammar $G' = (V', T, I', P', S')$ generating L, with the same index of G, such that $I' = I$ and the productions of P' are of the form:*

$$\textit{1) } A \to \nu, \quad \textit{2) } A \to Bf, \quad \textit{or} \quad \textit{3) } Af \to \nu,$$

where $A, B \in V'$, $f \in I'$ and $\nu \in (V' \cup T)^*$ is a word of the form

$$\nu = u, \quad \text{or} \quad \nu = uXZ, \quad \text{or} \quad \nu = uXv, \qquad X, Z \in V', \quad u, v \in T^*.$$

Proof of the Claim. Let us first assume that G has a sole production p of the form

$$A \to \nu = u_1 X_1 u_2 X_2 \cdots u_k X_k u_{k+1}, \quad k \geq 2,\ A, X_i \in V, u_i \in T^*. \tag{1}$$

Define the following list of productions:

i. $A \to u_1 X_1 Z_1$

ii. For every $j = 1, \ldots, k-2$, $\quad Z_j \to u_{j+1} X_{j+1} Z_{j+1}$

iii. $Z_{k-1} \to u_k X_k u_{k+1}$,

where $Z_j, (j = 1, \ldots, k-1)$, are new variables not in V.

Remove the production (1) from P, add to P the list of productions defined at (i)-(ii)-(iii) above, and add to V the corresponding list of new variables Z_j's. Let G' be the grammar obtained from G by using the previous transformation. We now observe that G' satisfies the claim and that the derivation of G defined by (1) is simulated by the derivation of G':

$$A \Rightarrow_{G'} u_1 X_1 Z_1 \Rightarrow_{G'} u_1 X_1 u_2 X_2 Z_2 \Rightarrow_{G'} \cdots \Rightarrow_{G'} u_1 X_1 u_2 \cdots u_{k-1} X_{k-1} Z_k \Rightarrow_{G'} \nu.$$

Moreover such derivation has index not larger than that of G. From the latter remark, it is easily checked, by induction on the length of the derivations of G', that G' has the same index of G and that $L = L(G')$.

The case of productions $Af \to \nu, f \in I$, is similarly treated. If G has two or more productions of the forms previously considered, the claim is obtained by iterating the previous argument.

Let $G = (V, I, T, P, S)$ be a finite-index indexed grammar in the form given by the previous Claim. Let R be a regular language over T and let $\mathcal{A} = (Q, T, \tau, q_0, K)$ be a finite deterministic and complete automaton accepting R, where Q is the set of states of $\mathcal{A}$, $\tau : Q \times T \to Q$ is its transition function, $q_0 \in Q$ is its unique initial state while K is the set of final states of $\mathcal{A}$. In the sequel, for the sake of simplicity, the extension of the function τ to the set $Q \times T^*$ will be still denoted by τ.
We proceed to construct a new finite-index grammar G' such that $G' = (V', I', T, P', S')$ and $L(G') = L \cap R$.

The set V' of variables of G' will be of the form $\langle p, X, q\rangle$, where p and q are in Q and X is in V, together with a new symbol S', denoting the start variable of G'.

The set I' of indices of G' is a copy of I disjoint with it. For every index f of I, we will denote by f' the corresponding copy of f in I' (it is understood that if $f = \lambda$ then $f' = \lambda$).

The set P' of productions of G' is defined as follows.

1. If $Af \to u$ is in P, where $f \in I \cup \{\lambda\}, u \in T^*$, and $\tau(p, u) = q$, then P' contains the set of productions $\langle p, A, q\rangle f' \to u$, for all $p, q \in Q$ such that p is transformed to q by reading u.

2. If $A \to Bf$ is in P, where $f \in I$, then P' contains the set of productions

$$\langle p, A, q\rangle \to \langle p, B, q\rangle f',$$

   where p, q are two arbitrary states of Q.

3. If $Af \to vDw$ is in P, where $f \in I \cup \{\lambda\}, A, D \in V, v, w \in T^*$, then P' contains, for all $p, q, r, s \in Q$, the set of productions

$$\langle p, A, q\rangle f' \to v\langle r, D, s\rangle w,$$

   provided that $\tau(p, v) = r$, and $\tau(s, w) = q$.

4. If $Af \to uBC$ is in P, where $f \in I \cup \{\lambda\}, A, B, C \in V, u \in T^*$, then P' contains, for all $p, q, r', r'' \in Q$, the set of productions

$$\langle p, A, q\rangle f' \to u\langle r', B, r''\rangle\langle r'', C, q\rangle,$$

   provided that $\tau(p, u) = r'$.

5. Finally P' contains the production $S' \to \langle s_0, S, p\rangle$, for all $p \in K$.

No other production different from the form specified in the list above is in P'.

The first task is to show that $L \cap R = L(G')$. For this purpose, we first show that: $\langle p, A, q\rangle f'_1 \cdots f'_i \Rightarrow^*_{G'} u$, with $i \geq 0$, $u \in T^*$, if and only if $Af_1 \cdots f_i \Rightarrow^*_G u$ and $\tau(p, u) = q$. Indeed, from this statement, we get $S' \Rightarrow_{G'} \langle s_0, S, q\rangle \Rightarrow^*_{G'} u$, for some $q \in K$, if and only if $S \Rightarrow^*_G u$, and $\tau(s_0, u) = q$, which is sufficient to complete the proof.

Let us first prove that:

(*) If $\langle p, A, q\rangle f'_1 \cdots f'_i \Rightarrow^\ell_{G'} u$ is a derivation of G' of length $\ell \geq 0$ then $Af_1 \cdots f_i \Rightarrow^*_G u$ and $\tau(p, u) = q$.

(*) is easily checked to be true for derivations of length 1. Now suppose that (*) is true for all $m < \ell$ with $m \geq 1$ and let $\langle p, A, q\rangle f'_1 \cdots f'_i \Rightarrow^\ell_{G'} u$ be a derivation of G' of length ℓ. Such a derivation can be of one of the following forms.

(i) $\langle p, A, q\rangle f'_1 \cdots f'_i \Rightarrow_{G'} \langle p, B, q\rangle f' f'_1 \cdots f'_i \Rightarrow^{\ell-1}_{G'} u, \quad f' \in I',$

that is, the first production of the derivation has the form (2). By the inductive hypothesis, we then have $Bff_1 \cdots f_i \Rightarrow^*_G u$ and $\tau(p, u) = q$, which yields $Af_1 \cdots f_i \Rightarrow_G Bff_1 \cdots f_i \Rightarrow^*_G u$ and $\tau(p, u) = q$.

(ii) $\langle p, A, q\rangle f' f'_1 \cdots f'_i \Rightarrow_{G'} v\langle r, D, s\rangle f'_1 \cdots f'_i w \Rightarrow^{\ell-1}_{G'} u$, $f' \in I' \cup \{\lambda\}$,

that is, the first production of the derivation has the form (3). Set $u = vu'w$. From the latter, we get $\langle r, D, s\rangle f'_1 \cdots f'_i \Rightarrow^{\ell-1}_{G'} u'$ so that, by the inductive hypothesis, $Df_1 \cdots f_i \Rightarrow^*_G u'$ and $\tau(r, u') = s$. On the other hand, we know that

$$Af \Rightarrow_G vDw, \qquad \tau(p, v) = r, \quad \tau(s, w) = q,$$

thus yielding $Aff_1 \cdots f_i \Rightarrow_G vDf_1 \cdots f_i w \Rightarrow^*_G vu'w = u$. Furthermore, $\tau(p, v) = r$, $\tau(s, w) = q$ which gives $\tau(p, u) = q$.

(iii) $\langle p, A, q\rangle f' f'_1 \cdots f'_i \Rightarrow_{G'} v\langle r', B, r''\rangle f'_1 \cdots f'_i \langle r'', C, q\rangle f'_1 \cdots f'_i \Rightarrow^{\ell-1}_{G'} u$,
$f' \in I' \cup \{\lambda\}, \quad r' = \tau(p, v)$,

that is, the first production of the derivation has the form (4). Set $u = vu'$, with $u' \in A^*$. From the second sentential form, we get

$$\langle r', B, r''\rangle f'_1 \cdots f'_i \Rightarrow^{\ell_1}_{G'} u'_1, \quad \langle r'', C, q\rangle f'_1 \cdots f'_i \Rightarrow^{\ell_2}_{G'} u'_2,$$

where $u' = u'_1 u'_2$, with $u'_1, u'_2 \in A^*, l_1 < l, l_2 < l$. By the inductive hypothesis, we have

$$Bf_1 \cdots f_i \Rightarrow^*_G u'_1, \quad Cf_1 \cdots f_i \Rightarrow^*_G u'_2,$$

together with

$$\tau(r', u'_1) = r'', \quad \tau(r'', u'_2) = q, \tag{2}$$

thus yielding

$$Aff_1 \cdots f_i \Rightarrow_G vBf_1 \cdots f_i Cf_1 \cdots f_i \Rightarrow^*_G vu'_1 Cf_1 \cdots f_i \Rightarrow^*_G vu'_1 u'_2 = vu' = u.$$

Finally, from (2) and $\tau(p, v) = r'$, we get $\tau(p, u) = q$.

(iv) $\langle p, A, q\rangle f' f'_1 \cdots f'_i \Rightarrow_{G'} wf'_1 \cdots f'_i \Rightarrow^{\ell-1}_{G'} u$,

that is, the first production of the derivation has the form (1). In this case, $f'_1 = \cdots = f'_i = \lambda$, and $\ell = 1$ so that the claim is trivially proved.

Since the latter cases represent all the possible ways an arbitrary derivation can start, (*) is proved. Similarly, taking into account the fact that the productions of G are in the form given in Claim 1, one proves by induction on the length of a derivation in G that if $Af_1 \cdots f_i \Rightarrow^{\ell}_G u$ is a derivation of G of length $\ell \geq 0$ and $\tau(p, u) = q$ then $\langle p, A, q\rangle f'_1 \cdots f'_i \Rightarrow^*_{G'} u$. By the previous remark, this implies that $L(G') = L(G) \cap R$.

Let δ' be a derivation of G'. By induction on the length of δ', one can prove the existence of a derivation δ of G that simulates (step by step) δ'. This implies that if G is a grammar of finite index, then G' is of the same type as well. This concludes the proof. □

Next, we show closure under an inverse morphism.

Let T and T' be two alphabets with $T \subseteq T'$ and let $\widehat{\pi_T} : (T')^* \to T^*$ be the projection of $(T')^*$ onto T^*, that is the epi-morphism from $(T')^*$ onto T^* generated by the mapping $\pi_T : T' \to T \cup \{\lambda\}$

$$\forall\ \sigma \in T', \pi_T(\sigma) = \begin{cases} \lambda & \text{if } \sigma \notin T, \\ \sigma & \text{if } \sigma \in T \end{cases}.$$

In the sequel, for the sake of simplicity, we denote the projection $\widehat{\pi_T}$ by π_T. It is useful to remark that, for every $w \in T^*$ and $w' \in (T')^*$, with $w = a_1 \cdots a_n$, $n \geq 0$, $a_i \in T$,

$$w' \in \pi_T^{-1}(w) \ \Leftrightarrow\ w' = w_1 a_1 \cdots w_n a_n w_{n+1},\ w_i \in (T' \setminus T)^*. \tag{3}$$

Lemma 13. *If $L \in \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ with $L \subseteq T^*$, then $\pi_T^{-1}(L) \in \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$.*

Proof. Let $G = (V, T, I, P, S)$ be a finite-index indexed grammar generating L. We construct a finite-index grammar G' generating $\pi_T^{-1}(L)$ with the same index.

For this purpose, let $p = Xf \to \nu$, with $X \in V, f \in I \cup \{\lambda\}$, and $\nu \in (VI^* \cup T^*)^*$, be a production of G of the form (1) or (3) (according to Definition 2). Then p has the form

$$Xf \to \nu = u_1 X_1 \cdots u_k X_k u_{k+1}, \quad u_i \in T^*,$$

where $X, X_i \in V$, with $i = 1, \ldots, k$, and, for every $i = 1, \ldots, k+1$,

$$u_i = a_{i,1} \cdots a_{i,n_i}, \quad n_i \geq 0, \quad a_{i,j} \in T.$$

Let us associate with p, the following set of productions:

- $Xf \to Y_{1,0} \cdots Y_{k,0}$,
- $\forall\, i = 1, \ldots, k, \forall\, j = 0, \ldots, n_i, \quad Y_{i,j} \to cY_{i,j}, \quad c \in T' \setminus T$,
- $\forall\, i = 1, \ldots, k, \forall\, j = 0, \ldots, n_i - 1, \quad Y_{i,j} \to a_{i,j+1} Y_{i,j+1}$,
- $\forall\, i = 1, \ldots, k-1, \quad Y_{i,n_i} \to X_i, \qquad Y_{k,n_k} \to Y'_{k,0}$,
- $\forall\, j = 0, \ldots, n_{k+1}, \quad Y'_{k,j} \to Y'_{k,j} c, \quad c \in T' \setminus T$,
- $\forall\, j = 0, \ldots, n_{k+1} - 1, \quad Y'_{k,j} \to Y'_{k,j+1} a_{k+1,n_{k+1}-j}$,
- $Y'_{k,n_k} \to X_k$,

where $Y_{i,j}$ and $Y'_{k,j}$ are new variables not in V.

Now remove the production p from P and add respectively to P and V the productions defined above and the corresponding set of new variables $Y_{i,j}$'s and $Y'_{k,j}$'s.

By applying the previous argument to every production p of the latter form, we will get a new grammar $G' = (V', T', I', P', S')$, where $I' = I$, $S' = S$ and the sets V' and P' are obtained from V and P respectively by iterating the latter combinatorial transformation.

It is useful now to remark that, in correspondence of every production $Xf \to u_1 X_1 \cdots u_k X_k u_{k+1}$, of G of the form (1) or (3), there exists a derivation of G' such that

$$Xf \Rightarrow_{G'}^* w_1 X_1 w_2 X_2 \cdots w_k X_k w_{k+1},$$

where, for all $i = 1, \ldots, k+1$, $w_i \in (T')^*$ and $w_i \in \pi_T^{-1}(u_i)$.

Taking into account the latter argument, the form of the new productions added to G', and Eq. (3), by induction on the length of the derivations of G and G' respectively, one proves the following two claims:

- for every $w' \in T'^*$, $S' \Rightarrow_{G'}^* w'$ if and only if there exists a derivation of G such that $S \Rightarrow_G^* w$, with $w \in T^*$, and $w' \in \pi_T^{-1}(w)$.
- if a non negative integer bounds the index of an arbitrary derivation of G the same does for G'. This implies that G' is a finite-index grammar with the same index of G.

This concludes the proof. □

Next, it is possible to show closure under rational transductions.

Lemma 14. *Let T and T' be two alphabets. Let $\tau : T^* \to (T')^*$ be a rational transduction from T^* into $(T')^*$. If L is a language of T^* in the family $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$, then $\tau(L) \in \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$.*

PROOF. Let us first assume that $T \cap T' = \emptyset$. By Nivat's theorem for the representation of rational transductions (see [4], Ch. III, Thm 4.1), there exists a regular set R over the alphabet $(T \cup T')$ such that

$$\tau = \{(\pi_T(u), \pi_{T'}(u)) : \ u \in R\},$$

where π_T and $\pi_{T'}$ are the projections of $(T \cup T')^*$ onto T^* and T'^* respectively.

From the latter, one has that, for every $u \in T^*$, $\tau(u) = \pi_{T'}(\pi_T^{-1}(u) \cap R)$, so that

$$\tau(L) = \bigcup_{u \in L} \tau(u) = \pi_{T'}(\pi_T^{-1}(L) \cap R). \tag{4}$$

Since, by hypothesis, $L \in \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$, the claim follows from (4), by applying Lemma 11, 12, and 13.

Let us finally treat the case where T and T' are not disjoint. Let T'' be a copy of T' with $T'' \cap T = \emptyset$ and let $c_{T''} : (T')^* \to (T'')^*$ be the corresponding copying iso-morphism from $(T')^*$ onto $(T'')^*$. Since $T'' \cap T = \emptyset$, by applying the latter argument to the rational transduction $c_{T''}\tau : T^* \to (T'')^*$, one has $(c_{T''}\tau)(L) \in \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$. Since $c_{T''}^{-1}((c_{T''}\tau)(L)) = \tau(L)$, then the claim follows from the latter by applying Lemma 11. □

Since inverse morphisms are rational transductions, the following is immediate:

Corollary 15. *$\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ is closed under inverse morphisms.*

By Lemma 11, Lemma 12, and Corollary 15, we obtain:

Theorem 16. *The family $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ is a full trio.*

We now prove a result which extends the semi-linearity of a family of languages to a bigger family. If $\mathcal{C}$ is a full trio of semi-linear languages and $\mathcal{L}$ is the family of languages $\mathcal{L}(\mathsf{NCM})$ accepted by NCMs, let $\mathcal{C} \wedge \mathcal{L} = \{L_1 \cap L_2 : L_1 \in \mathcal{C}, L_2 \in \mathcal{L}\}$.

Theorem 17. *Let $\mathcal{C}$ be a full trio of semi-linear languages. Every language in $\mathcal{C} \wedge \mathcal{L}(\mathsf{NCM})$ has a semi-linear Parikh image.*

PROOF. Let A and B be disjoint alphabets. Consider the homomorphism

$$\widehat{\pi_A} : (A \cup B)^* \to A^*$$

defined before Lemma 13. If L is a language over A^*, then $\widehat{\pi_A}^{-1}(L) = \{x : x \in (A \cup B)^*, h(x) \in L\}$.

Let $A = \{a_1, \ldots, a_n\}$ and $L_1 \subseteq A^*$ be in $\mathcal{C}$. Then $\widehat{\pi_A}^{-1}(L_1)$ is also in $\mathcal{C}$, since $\mathcal{C}$ is closed under inverse homomorphism. Note that the Parikh image of L_1, $\psi(L_1)$, is semi-linear since $\mathcal{C}$ is a semi-linear family.

Now let $L_2 \subseteq A^*$ be a language accepted by an NCM M_2. Any NCM can be simulated by an NCM M_2 whose counters are 1-reversal [3]. We may assume that a string is accepted by M_2 if and only if it enters a unique halting state f with all counters zero.

Let M_2 have k 1-reversal counters. Let $B = \{p_1, q_1, \ldots, p_k, q_k\}$ be new symbols disjoint from A. Construct an NFA M_3 which when given a string w in $(A \cup B)^*$ simulates M_2, but whenever counter c_i increments, M_3 reads the next input symbol and checks that it is p_i. When M_2 decrements counter c_i, M_3 reads q_i from the input. (Note that after the first q_i is read, no p_i should appear on the remaining input symbols.) M_3 guesses when each counter c_i becomes zero (this may be different time for each i), after which, M_3 should no longer read q_i. At some point, M_3 guesses that all counters are zero. It continues the simulation and when M_2 accepts in state f, M_3 accepts. Clearly, a string x in A^* is accepted by M_2 if and only if there is a string w in $(A \cup B)^*$ accepted by M_3 such that:

(1) $\widehat{\pi_A}(w) = x$,
(2) $|w|_{p_i} = |w|_{q_i}$ for each $1 \leq i \leq k$.

Let R_3 be the regular set accepted by M_3. Since $\mathcal{C}$ is a full trio:

$$\widehat{\pi_A}^{-1}(L_1) \in \mathcal{C}, \quad L_4 = (\widehat{\pi_A}^{-1}(L_1) \cap R_3) \in \mathcal{C}.$$

Hence the Parikh image of $L_4, \psi(L_4)$, is a semi-linear set Q_4.

Now $A = \{a_1, \ldots, a_n\}$ and $B = \{p_1, q_1, \ldots, p_k, q_k\}$. Define the semi-linear set

$$Q_5 = \{(s_1, \ldots, s_n, t_1, t_1, \ldots, t_k, t_k) : s_i, t_i \geq 0\}.$$

(Note that the first n coordinates refer to the counts corresponding to symbols $a_1, \ldots, a_n$, and the last $2k$ coordinates refer to the counts corresponding to symbols $(p_1, q_1, \ldots, p_k, q_k)$.)

Then $Q_6 = Q_4 \cap Q_5$ is semi-linear, since semi-linear sets are closed under intersection. Now $\psi(L_1 \cap L_2)$ coincides with the projection of Q_6 on the first n coordinates. Hence $\psi(L_1 \cap L_2)$ is semi-linear, since semi-linear sets are closed under projections. □

Note that the above proposition does not depend on how the languages in $\mathcal{C}$ are specified. It extends the semi-linearity of languages in $\mathcal{C}$ to a bigger family that can do some "counting". The theorem applies to all well-known full trios of semi-linear languages, in particular, to $\mathcal{C} = \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$.

Corollary 18. *Let $\mathcal{C}$ be a full trio whose closures under homomorphism, inverse homomorphism and intersection with regular sets are effective. Moreover, assume that for each L in $\mathcal{C}$, $\psi(L)$ can effectively be constructed. Then $\mathcal{C} \wedge \mathcal{L}(\mathsf{NCM})$ has a decidable emptiness problem.*

Indeed, decidability of emptiness follows immediately from effective construction of the semilinear set [13] as having any vector describe a linear set implies the language is non-empty, and no vector implies the language is empty. Note that $\mathcal{L}(\mathsf{NCM})$ is also a full trio of semi-linear languages. It is easy to see that the theorem is not true if $\mathcal{L}(\mathsf{NCM})$is replaced with a different full trio. For example suppose $\mathcal{C} = \mathcal{L}$ is the family of languages accepted by 1-reversal NPDAs (= linear context-free languages). Let

$$L_1 = \{a^{n_1}\sharp \cdots \sharp a^{n_k}\$a^{n_k}\sharp \cdots \sharp a^{n_1} : k \geq 4, n_i \geq 1\},$$

$$L_2 = \{a^{n_1}\sharp \cdots \sharp a^{n_k}\$a^{m_k}\sharp \cdots \sharp a^{m_1} : k \geq 4,\ n_i,\ m_i \geq 1,\ m_j = n_{j+1},\ 1 \leq j < k\}.$$

Clearly, L_1 and L_2 can be accepted by 1-reversal NPDAs. But $L_1 \cap L_2$ is $\{(a^n\sharp)^{k-1}a^n\$(a^n\sharp)^{k-1}a^n : n \geq 1,\ k \geq 4\}$ and it is not semi-linear.

Similarly, it is known that the theorem does not hold when $\mathcal{C} = \mathcal{L}$ is the family of languages accepted by NFAs with one unrestricted counter (i.e., NPDAs with a unary stack alphabet in addition to a distinct bottom of the stack symbol which is never altered), as similar languages L_1, L_2 in this family can be constructed such that their intersection is not semilinear (Proposition 31 of [7]).

4. Bounded Languages and Hierarchy Results

The purpose of this section is to demonstrate that all bounded Ginsburg semi-linear languages are in $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ (thus implying they are in $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ as well), but not in $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$.

Notice that the language L from the remarks following Theorem 9 is a bounded Ginsburg semi-linear language. Thus, the following is true:

Theorem 19. *There are bounded Ginsburg semi-linear languages that are not in $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$.*

Furthermore, it has been shown that in every semi-linear full trio, all bounded languages in the family are bounded Ginsburg semi-linear [14]. Further, $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ is a semi-linear full trio [6]. Therefore, the bounded languages in $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ are strictly contained in the bounded languages contained in any family containing all bounded Ginsburg semi-linear languages. We only mention here three of the many such families mentioned in [14].

Corollary 20. *The bounded languages in* $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ *are strictly contained in the bounded languages from* $\mathcal{L}(\mathsf{NCM}), \mathcal{L}(\mathsf{DCM}), \mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}})$.

Theorem 21. $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ *contains all bounded Ginsburg semi-linear languages.*

Proof. We now prove that if L is a bounded Ginsburg semi-linear language, with $L \subseteq u_1^* \cdots u_k^*$, then $L \in \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$. By Definition 1, $L = \varphi(B)$, where B is a semi-linear subset of $\mathbb{N}^k$. Since $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ is closed under union by Lemma 10, it is enough to show it for a linear set B. Let B be a set of the form $B = \{\mathbf{b}_0 + x_1\mathbf{b}_1 + \cdots + x_\ell \mathbf{b}_\ell \ : x_1, \ldots, x_\ell \in \mathbb{N}\}$, where $\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_\ell$, are vectors of $\mathbb{N}^k$. By denoting the arbitrary vector $\mathbf{b}_i$ as $(b_{i1}, \ldots, b_{ik})$, we write B as

$$\{(b_{01} + x_1 b_{11} + \cdots + x_\ell b_{\ell 1},\ \ldots,\ b_{0k} + x_1 b_{1k} + \cdots + x_\ell b_{\ell k}),\ : x_1, \ldots, x_\ell \in \mathbb{N}\},$$

so that the language $L = \varphi(B)$ becomes

$$u_1^{b_{01}+x_1 b_{11}+\cdots+x_\ell b_{\ell 1}} u_2^{b_{02}+x_1 b_{12}+\cdots+x_\ell b_{\ell 2}} \cdots u_k^{b_{0k}+x_1 b_{1k}+\cdots+x_\ell b_{\ell k}}, \tag{5}$$

where $x_1, \ldots, x_\ell \in \mathbb{N}$. Let us now define an indexed grammar G such that $L = L(G)$. Let $G = (V, T, I, P, S)$, where

$$V = \{S, Y, X_1, \ldots, X_k\}, \quad T = A, \quad I = \{e, f_1, f_2, \ldots, f_\ell\},$$

and the set P of productions is the following:

1. $P_{start} = (S \to Ye)$
2. For every $j = 1, \ldots, \ell$, $P_j = (Y \to Yf_j)$
3. $Q = (Y \to X_1 X_2 \cdots X_k)$
4. For every $i = 1, \ldots, k$ and for every $j = 1, \ldots, \ell$,

$$R_{i0} = (X_i e \to u_i^{b_{0i}}), \quad R_{ij} = (X_i f_j \to u_i^{b_{ji}} X_i).$$

Let us finally prove that $L = L(G)$ and G is an uncontrolled grammar. Let us first show that $L \subseteq L(G)$. Let $w \in L$. By (5), there exist $x_1, \ldots, x_\ell \in \mathbb{N}$ such that

$$w = u_1^{b_{01}+x_1 b_{11}+\cdots+x_\ell b_{\ell 1}} u_2^{b_{02}+x_1 b_{12}+\cdots+x_\ell b_{\ell 2}} \cdots u_k^{b_{0k}+x_1 b_{1k}+\cdots+x_\ell b_{\ell k}}.$$

Consider the derivation defined by the word over the alphabet P:

$$\mathcal{P} = P_{start} P_1^{x_1} P_2^{x_2} \cdots P_\ell^{x_\ell} Q Q_1 \cdots Q_k,$$

where, for every $i = 1, \ldots, k$, $Q_i = R_{i\ell}^{x_\ell} \cdots R_{i2}^{x_2} R_{i1}^{x_1} R_{i0}$. It is easily checked that $S \Rightarrow_{\mathcal{P}} w$. Indeed,
$S \Rightarrow_{P_{start}} Ye \Rightarrow_{P_1^{x_1} P_2^{x_2} \cdots P_\ell^{x_\ell}} Yf_\ell^{x_\ell} \cdots f_1^{x_1} e \Rightarrow_Q$
$X_1 f_\ell^{x_\ell} \cdots f_1^{x_1} e \cdots X_k f_\ell^{x_\ell} \cdots f_1^{x_1} e \Rightarrow_{Q_1} u_1^{b_{01}+x_1 b_{11}+\cdots+x_\ell b_{\ell 1}} X_2 \cdots X_k f_\ell^{x_\ell} \cdots f_1^{x_1} e$
$\Rightarrow_{Q_2} u_1^{b_{01}+x_1 b_{11}+\cdots+x_\ell b_{\ell 1}} u_2^{b_{02}+x_1 b_{12}+\cdots+x_\ell b_{\ell 2}} X_3 \cdots X_k f_\ell^{x_\ell} \cdots f_1^{x_1} e \Rightarrow_{Q_3 \cdots Q_k} w,$

so that $w \in L(G)$. Similarly, it can be shown that $L(G) \subseteq L$. Thus $L = L(G)$. Moreover, taking into account the form of the productions of G, it is easily checked that the index of every derivation of G is not larger than k. □

Since it is known that in any semi-linear full trio, all bounded languages in the family are bounded Ginsburg semi-linear, the bounded languages in $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ coincide with several other families, including a deterministic machine model [14].

Corollary 22. *The bounded languages in* $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ *coincide with the following families of languages: bounded Ginsburg semi-linear languages, bounded languages in* $\mathcal{L}(\mathsf{NCM}), \mathcal{L}(\mathsf{DCM})$, $\mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}})$*, the class of string languages of simple (i.e., linear and non deleting) tree grammars (see [15]) and several other families listed in [14].*

Also, since $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ does not contain all bounded Ginsburg semi-linear languages by Theorem 19, but $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ does, the following is immediate:

Corollary 23. *The bounded languages in* $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ *are strictly contained in the bounded languages of* $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$.

Next, a restriction of $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ is studied and compared to the other families. And indeed, this family is quite general as it contains all bounded Ginsburg semi-linear languages in addition to some languages that are not in $\mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}})$.

Now let $p = (Af \to \nu) \in P$, with $f \in I \cup \{\lambda\}$, be a production. Then p is called *special* if the number of occurrences of variables of V in ν is at least 2, and *linear*, otherwise. Denote by $P_{\mathcal{S}}$ and $P_{\mathcal{L}}$ the sets of special and linear productions of P respectively. By Definition 6, a grammar G is uncontrolled finite-index if and only if the number of times special productions appear in every successful derivation of G is upper bounded by a given fixed integer (not depending on the derivation).

Next, we will deal with uncontrolled grammars such that in every successful derivation of G, at most one special production occurs. The languages generated by such grammars form a family denoted $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$. It is worth noticing that a careful rereading of the proof of Theorem 16 and Lemma 10 shows that they hold for $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$ as well. Further, it is clear that only one special production is used in every derivation of a word in the proof of Theorem 21. Therefore, the following holds:

Theorem 24. *The family* $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$ *is a union-closed full trio and it contains all bounded Ginsburg semi-linear languages.*

It is immediate from the definitions that $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}}) \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1}) \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$. Further, since $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$ contains all bounded Ginsburg semi-linear languages by Theorem 24, but the linear indexed languages do not, by Theorem 19, the following holds:

Theorem 25. $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}}) \subset \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1}) \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$.

Then the following is true from [14].

Corollary 26. $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$ *is a semi-linear full trio containing all bounded Ginsburg semi-linear languages. Further, the bounded languages in* $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$, $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$, $\mathcal{L}(\mathsf{NCM})$, $\mathcal{L}(\mathsf{DCM})$, *and* $\mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}})$ *all coincide, (also with several others listed in [14]).*

5. Some Examples, Separation, and Decidability Results

We start this section by giving an example that clarifies previous results.

Example 1. *Let* $L = \{a^n b^n c^n \$ a^n b^n c^n : n \in \mathbb{N}\}$. *If* $\varphi : \mathbb{N}^7 \to a^*b^*c^*\$^*a^*b^*c^*$, *then* $L = \varphi(B)$, *where* $B = \{\mathbf{b}_0 + n\mathbf{b}_1 : n \in \mathbb{N}\}$, *with* $\mathbf{b}_0 = (0,0,0,1,0,0,0)$ *and* $\mathbf{b}_1 = (1,1,1,0,1,1,1)$. *It is worth noticing that, by the discussion preceding Theorem 19, L is not a linear indexed language. We define an uncontrolled finite-index grammar* $G = (V, T, I, P, S)$ *where* $V = \{S, Y, X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$, $T = \{a, b, c, \$\}$, $I = \{e, f\}$, *and the set* P *of productions is:*

$$P_{start} = S \to Ye,\ P = Y \to Yf,\ Q = Y \to X_1X_2\cdots X_7$$

$$\begin{array}{lllllll} X_1f \to aX_1 & X_2f \to bX_2 & X_3f \to cX_3 & X_4f \to X_4 & X_5f \to aX_5 & X_6f \to bX_6 \\ X_7f \to cX_7 & X_1e \to \lambda & X_2e \to \lambda & X_3e \to \lambda & X_4e \to \$ & X_5e \to \lambda \\ X_6e \to \lambda & X_7e \to \lambda. \end{array}$$

For an arbitrary derivation, we get

$$S \Rightarrow Ye \Rightarrow^n Yf^ne = X_1f^neX_2f^neX_3f^neX_4f^neX_5f^neX_6f^neX_7f^ne \Rightarrow^* a^nb^nc^n\$a^nb^nc^n.$$

As the only freedom in derivations of G consists of how many times the rule P is applied and of trivial variations in order to perform the rules $X_if \to \sigma X_i, \sigma \in T \cup \{\varepsilon\}$, it should be clear that $L = L(G)$.

It is known that decidability of several properties holds for semi-linear trios where the properties are effective [13]. This is the case for $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$, and also for $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ [6].

Corollary 27. *Containment, equality, membership, and emptiness are decidable for bounded languages in $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ and $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$.*

Lastly, it is known that $\mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}})$ cannot generate some context-free languages [19], but all context-free languages can be generated by indexed linear grammars by Theorem 8, which are all in $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$.

Corollary 28. *There are languages in $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1})$ and $\mathcal{L}(\mathsf{IND}_{\mathsf{LIN}})$ that are not in $\mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}})$.*

We provide an example of a language in $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ whose Parikh image is not a semi-linear set.

Example 2. *We construct a grammar of index 3, which is not uncontrolled, that generates the language $L = \{aba^2b \cdots a^nba^{n+1} \ : \ n \geq 1\}$. Let $G = (V, T, I, P, S)$ be the grammar where $V = \{S, A, B, X, X', X''\}$, $T = \{a, b\}$, $I = \{e, f, g\}$, and the set of productions of G are defined as:*

- $p_0 = S \to Xe, \quad p_1 = X \to AfBfX'f, \quad p_2 = X' \to X, \quad p_3 = X' \to X'',$
- $p_4 = X''f \to aX'', \quad p_5 = X''e \to a, \quad p_6 = Af \to aA, \quad p_7 = Ae \to \lambda,$
- $p_8 = Bf \to B, \quad p_9 = Be \to b.$

One can check that G satisfies the properties mentioned above.

Let G' be the grammar obtained from G by replacing the production p_9 above with $(Be \to \lambda)$. Then one verifies that G' is a grammar of index 3 generating the unary language $\{a^{n(n+1)/2} \ : \ n \geq 2\}$, that is not bounded Ginsburg semi-linear.

From Example 2 we get

Corollary 29. *There are languages in $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ that are not semi-linear. Furthermore, there are bounded (and unary) languages in $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ that are not bounded Ginsburg semi-linear.*

This allows for the separation of $\mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}})$ (which only contains semi-linear languages) and $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$.

Corollary 30. $\mathsf{CFL} \subset \mathcal{L}(\mathsf{IND}_{\mathsf{LIN}}) \subset \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}_1}) \subseteq \mathcal{L}(\mathsf{IND}_{\mathsf{UFIN}}) \subset \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$.

Finally, we show that all finite-index $\mathsf{ET0L}$ languages are finite-index indexed languages.

Theorem 31. $\mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}}) \subset \mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$.

Proof. Strictness follows since $\mathcal{L}(\mathsf{IND}_{\mathsf{FIN}})$ contains non-semi-linear languages by Corollary 29, however $\mathcal{L}(\mathsf{ET0L}_{\mathsf{FIN}})$ only contains semi-linear languages [17].

We refer to [17] for the formal definitions of $\mathsf{ET0L}$ systems and finite-index $\mathsf{ET0L}$ systems, which we will omit.

Let $G = (V, \mathcal{P}, S, T)$ be a k-index $\mathsf{ET0L}$ system. We can assume without loss of generality that G is in so-called active-normal form, so that the set of active symbols of V (those that can be changed by some production table) is equal to $V \setminus T$. Let $\mathcal{P} = \{f_1, \ldots, f_r\}$ be the set of production tables. Then create an indexed grammar $G' = (V', T, I, P, S')$ where $V' = (V \setminus T) \cup \{S'\}$, S' is a new variable, $I = \{f_1, \ldots, f_r\}$, and P contains the following productions:

1. $S' \to S' f_i, \forall i, 1 \leq i \leq r$,

2. $S' \to S$,

3. $B f_i \to \nu, \forall (B \to \nu) \in f_i, B \in V \setminus T$.

Let $w \in L(G)$. Then $w_0 \Rightarrow_{f_{j_1}} w_1 \Rightarrow \cdots \Rightarrow_{f_{j_l}} w_l, w_0 = S, w_l = w$. Let w'_i be obtained from w_i by placing $f_{j_{i+1}} \cdots f_{j_l}$ after each variable of w_i.

We will show by induction on i, $0 \leq i \leq l$, that $S' \Rightarrow^*_{G'} w'_i$. Indeed, $S' \Rightarrow^*_{G'} S f_{j_1} \cdots f_{j_l} = w'_0$, by using productions of type 1 followed by 2. Assume the inductive hypothesis is true for some i, $0 \leq i < l$. Then $S' \Rightarrow^*_{G'} w'_i$. Then the next index on every variable of w'_i is $f_{j_{i+1}}$. Applying the productions corresponding to those used in the derivation $w_i \Rightarrow_{f_{j_{i+1}}} w_{i+1}$ in table $f_{j_{i+1}}$ on each variable of w'_i one at a time from left-to-right created in 3. of the construction above, w'_{i+1} is obtained. It is also clear that if the original derivation is of index-k, then the resulting derivation is of index-$2k$ (since the derivation of the indexed grammar proceeds sequentially instead of in parallel, the number of variables of the indexed grammar could potentially be more than k, but it is always less than the number of variables in the sentential form of the ET0L system plus the next sentential form).

Let $w \in L(G')$. Thus, $w_0 \Rightarrow_{p_1} w_1 \Rightarrow_{p_2} \cdots \Rightarrow_{p_l} w_l$, where $S' = w_0$ and $w_l = w \in T^*$. It should also be clear that we can assume without loss of generality that this derivation proceeds by rewriting variables in a "sweeping left-to-right" manner. That is, if $w_i = w'_i B w''_i$ derives w_{i+1} by rewriting variable B, then w_{i+1} derives w_{i+2} by rewriting the first variable of w''_i if it exists, and if not, the first variable of w_{i+1}. Then one "sweep" of the variables by rewriting each variable is similar to one rewriting step of an ET0L system. This is akin to a breadth first traversal on the derivation tree of w.

By the construction, there exists $\alpha > 0$ such that $p_1, \ldots, p_\alpha$ are productions created in step 1, $p_{\alpha+1}$ is created in step 2, and $p_{\alpha+2}, \ldots, p_l$ are created in step 3. Let $\beta_1, \ldots, \beta_q$ be such that $\beta_1 = \alpha + 2$, and the derivation from w_{β_i} is the start of the ith "sweep" from left-to-right, and let $\beta_{q+1} = l$. For $1 \leq i \leq q+1$, let u_i be obtained from w_{β_i} by removing all indices (so $u_{q+1} = w_l$).

We will show by induction that for all i, $1 \leq i \leq q+1$, it is true that $S \Rightarrow^*_G u_i$, and all variables in w_{β_i} are followed by the same index sequence. Indeed, $w_{\beta_1} = w_{\alpha+2} = S\gamma$ for some $\gamma \in I^*$, $u_1 = S$, and $S \Rightarrow^*_G u_1 = S$. Assume that the inductive hypothesis holds for some i, $1 \leq i \leq q$. Then in w_{β_i}, all variables are followed by the same index sequence. Let f be the first index following every variable. Then in the subderivation $w_{\beta_i} \Rightarrow_{p_{\beta_i}} \cdots \Rightarrow_{p_{\beta_{i+1}}} w_{\beta_{i+1}}$, because all productions applied were created in step 3, they must all pop the first index, and since they all start with the same index, they must all have been created from productions in the same table f. It is clear that $u_i \Rightarrow_G u_{i+1}$ using production table f. It is also immediate that all variables in $w_{\beta_{i+1}}$ are followed by the same sequence of indices. The proof follows. □

It is an open question though as to how $\mathcal{L}(\mathsf{ET0L_{FIN}})$ compares to $\mathcal{L}(\mathsf{IND_{UFIN}})$. For finite-index ET0L, uncontrolled systems, defined similarly to our definition of uncontrolled, is identical to finite-index ET0L. Furthermore, it is known that $\mathcal{L}(\mathsf{ET0L_{FIN}})$ is closed under Kleene-$*$ [17] and therefore contains $\{a^n b^n c^n : n > 0\}^*$. But we conjecture that this language is not in $\mathcal{L}(\mathsf{IND_{UFIN}})$ despite being in $\mathcal{L}(\mathsf{IND_{FIN}})$ by the proposition above. This would imply that $\mathcal{L}(\mathsf{IND_{UFIN}})$ is incomparable with $\mathcal{L}(\mathsf{ET0L_{FIN}})$ by Corollary 28.

Moreover it would be interesting to know whether the inclusion $\mathcal{L}(\mathsf{IND_{UFIN_1}}) \subseteq \mathcal{L}(\mathsf{IND_{UFIN}})$ is strict or not. The examples presented in this paper would suggest that the two latter families could be equal.

We finally note that the class of linear indexed languages studied by Duske and Parchmann is a proper subset of the one studied by Gazdar and Vijayy-Shanker. An example of a language in the second class but not in the first, is the language $L = \{a^n b^n c^n \$ a^m b^m c^m \mid n, m \geq 0\}$ which appears in the remarks following Theorem 9. It might then also be interesting to study the class of Gazdar and Vijayy-Shanker in connection with finite index restrictions.

References

[1] A. V. Aho, Indexed grammars—an extension of context-free grammars, J. ACM, 15 (4), 647–671, (1968).

[2] A. V. Aho, Nested stack automata, J. ACM, 16, 383–406, (1969).

[3] B. S. Baker, R. V. Book, Reversal-Bounded Multipushdown Machines, J. Comput. Syst. Sci., 8, 315–332, (1974).
[4] J. Berstel, *Transductions and Context-Free Languages*, B.B. Teubner, Stuttgart, 1979.
[5] J. Dassow, Gh. Păun, *Regulated Rewriting in Formal Language Theory,* EATCS Monographs on Theoretical Computer Science, 18, Springer-Verlag, Berlin, 1989.
[6] J. Duske, R. Parchmann, Linear indexed grammars, Theoret. Comput. Sci. 32, 47–60, (1984).
[7] J. Eremondi, O. H. Ibarra, and I. McQuillan, On the Complexity and Decidability of Some Problems Involving Shuffle, Information and Computation, 259 (2), 214–224, (2018).
[8] G. Gazdar, Applicability of Indexed Grammars to Natural Languages, pp. 69–94, Springer Netherlands, Dordrecht, (1988).
[9] S. Ginsburg, *The Mathematical Theory of Context-free Languages*, Mc Graw- Hill, New York, 1966.
[10] M. A. Harrison, *Introduction to Formal Language Theory*, Addison-Wesley Publishing Co., Reading, Mass., 1978.
[11] J. E. Hopcroft, J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley Publishing Co., Reading, Mass., 1979.
[12] O. H. Ibarra, Reversal-bounded multicounter machines and their decision problems, J. ACM, 1 (25), 116–133, (1978).
[13] O.H. Ibarra and I. McQuillan, On bounded semilinear languages, counter machines, and finite-index ET0L. In: Y. Han and K. Salomaa (eds.), Lecture Notes in Computer Science, 21st International Conference on Implementation and Application of Automata, CIAA 2016, Seoul, South Korea, vol. 9705, pp. 138–149, (2016).
[14] O.H. Ibarra and I. McQuillan, On families of full trios containing counter machine languages. In: S. Brlek and C. Reutenauer (eds.), Lecture Notes in Computer Science, 20th International Conference on Developments in Language Theory, DLT 2016, Montreal, Canada, vol. 9840, pp. 216–228, (2016).
[15] M. Kanazawa, A Generalization of Linear Indexed Grammars Equivalent to Simple Context-Free Tree Grammars. In: G. Morrill, R. Muskens, R. Osswald, F. Richter (eds.), Lecture Notes in Computer Science, 19th International Conference, FG 2014, Tübingen, Germany, 2014, vol. 8612, pp. 86–103, (2014).
[16] G. Rozenberg and A. Salomaa, *The Mathematical Theory of L Systems*, Academic Press, Inc., New York, 1980.
[17] G. Rozenberg and D. Vermeir, On ET0L systems of finite index, Information and Control, 38, 103–133, (1978).
[18] G. Rozenberg and D. Vermeir, On the effect of the finite index restriction on several families of grammars, Information and Control, 39, 284–302, (1978).
[19] B. Rozoy, The Dyck language $D_1'^*$ is not generated by any matrix grammar of finite index, Information and Computation 74 (1), 64–89, (1987).
[20] K. Vijay-Shanker and D. J. Weir, The equivalence of four extensions of context-free grammars, Mathematical Systems Theory, 27 (6), 511–546, (1994).
[21] G. Zetzsche, An Approach to Computing Downward Closures. In: M.M. Halldórsson and K. Iwama and N. Kobayashi and B. Speckmann (eds.), Automata, Languages, and Programming: 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part II, pp. 440–451, (2015).