Homomorphic Characterizations of Indexed Languages

Severine Fratani, El Makki Voundy

Homomorphic Characterizations of Indexed Languages

Séverine Fratani and El Makki Voundy

Aix-Marseille Université, CNRS, LIF UMR 7279, 13000, Marseille, France

Abstract. We study a family of context-free languages that reduce to ε in the free group and give several homomorphic characterizations of indexed languages relevant to that family.

Keywords: grammars, homomorphic characterizations, transductions, indexed languages

1 Introduction

The well known Chomsky–Schützenberger theorem [6] states that every context-free language L can be represented as $L = h(R \cap \mathcal{D}_k)$, for some integer k, regular set R and homomorphism h. The set $\mathcal{D}_k$ used in this expression, called Dyck language, is the set of well-bracketed words over k pairs of brackets. Combined with Nivat's characterization of rational transductions, this means that any context-free language can be defined as a set $L = h(g^{-1}(\mathcal{D}_2) \cap R)$, for some regular set R, and homomorphisms h and g.

Let us consider wider families of languages, Maslov defines in [13] an infinite hierarchy of languages included in recursively enumerable languages. The level 1 consists of context-free languages, the level 2 of indexed languages (initially defined by Aho [1]). Known as higher order languages since the last decades, the languages of the hierachy and derived objects as higher order trees [12], higher order schemes [9], or higher order graphs [5], are used to model programming languages and are in the core of the recent researches in program verification[18].

It is stated in [14] and proved in [8] that each level $\mathcal{L}_k$ of the hierarchy is a principal rational cone generated by a language $M_k \in \mathcal{L}_k$. This means that each language in $\mathcal{L}_k$ is the image of M_k by a rational transduction. Roughly speaking, the language M_k consists of words composed by k embedded Dyck words and can be viewed as a generalization of the Dyck language. Indeed it gives a description of derivations of an indexed grammar of level k, in the same way that the Dyck language encodes derivations of a context-free grammar.

This latter characterization describes $\mathcal{L}_k$ from a single language M_k, but this one is very complicated as soon as $k \geq 2$, as the majority of higher order languages. To better understand higher order languages, we think that it is necessary to characterize them using more simple objects. So, we may wonder whether it is possible to give versions of the Chomsky–Schützenberger theorem and a characterization by transduction of the level $k+1$ of the hierarchy, using

only the level k of the hierarchy. The fundamental point is then to identify mechanisms that bridge the level k to the level $k+1$.

In this paper, we solve the problem for the class IL of Indexed Languages (the level 2 of the hierarchy). In order to localize the problem, let us remark that from [10], recursively enumerable languages are sets that can be written as $L = h(K \cap \mathcal{D}_k)$ where K is a context-free language, and h a homomorphism. So if we want a homomorphic characterization of IL using only context-free or regular languages, we would have to consider a restricted class of context-free languages.

For this purpose, we introduce the class of ε-Reducible Context-Free Languages (ε-CFLs), which is a strict subclass of context-free languages that reduce to ε by the bilateral reduction $S = \{a\bar{a} \to \varepsilon, \bar{a}a \to \varepsilon\}_{a\in\Gamma}$ (these languages are thus defined over an alphabet Γ and its copy $\bar{\Gamma}$). We extend this definition to transductions: an ε-Reducible Context-Free Transduction (ε-CFT) is a context-free transduction whose domain is an ε-CFL. Using these objects, we obtain simple and elegant generalizations of the Chomsky–Schützenberger theorem. Indexed languages are:

- the images of $\mathcal{D}_2$ by ε-reducible context-free transductions. (Theorem 15);
- sets $h(Z \cap \mathcal{D}_k)$; where k is an integer, Z an ε-CFL, and h a homomorphism (Theorem 18).

Beyond these two results, we study the classes of ε-CFLs and ε-CFTs defined by means of context-free grammars and context-free transduction grammars. First we express them using *symmetric homomorphisms* which are homomorphisms under which there are closed. We establish a Chomsky–Schützenberger-like Theorem for ε-CFLs, and a Nivat-like characterization for ε-CFTs: every ε-CFL L can be represented as $L = g(R \cap \mathcal{D}_k)$ for some integer k, regular language R, and symmetric homomorphism g; and ε-CFTs are relations that can be represented as $\{(g(x), h(x)) \mid x \in R \cap \mathcal{D}_k\}$ for some integer k, regular language R, homomorphism h and symmetric homomorphism g. This leads to a third characterization: indexed languages are languages that can be described as $L = h(g^{-1}(\mathcal{D}_2) \cap R \cap \mathcal{D}_k)$, for some integer k, regular language R, homomorphism h and symmetric homomorphism g (Corollary 17).

Similar characterizations have been given for subclasses of indexed languages, by Weir [20] for linear indexed languages, by Kanazawa [11] and Sorokin [19] for yields of tree languages generated by simple context-free grammars. The main difference is that in their cases, the homomorphism g is not symmetric, but is *fixed* in function of k.

Overview. Section 1 is devoted to the study of ε-CFLs. After introducing necessary notions as free groups and Dyck languages, we define the class of grammars generating ε-CFLs. We then study their closure properties, and conclude the section by giving a Chomsky–Schützenberger-like characterization of the class of ε-CFLs. In Section 2, we extend our definition to transductions and define the class of ε-CFTs. After a subsection giving background on transductions we

give a Nivat-like characterization of ε-CFTs. The last section is devoted to indexed languages. After introducing indexed grammars, we prove that indexed languages are images of the Dyck language by ε-CFTs and deduce from this result several homomorphic characterizations.

2 Epsilon-Reducible Context-Free Languages

In this section, we study a family of context-free languages defined over a union of an alphabet and its opposite-disjoint copy and that reduce to the neutral element ε when projected into the free group. The main result here is a Chomsky–Schützenberger-like homomorphic characterization of these languages. We assume the reader to be familiar with context-free grammars and languages (see [3] for example), and present below a few necessary notions on free groups.

2.1 Free groups and Dyck languages

Given an alphabet Γ, we denote by $\overline{\Gamma}$ a disjoint copy $\overline{\Gamma} = \{\bar{a} \mid a \in \Gamma\}$ of it, and by $\widehat{\Gamma}$ the set $\Gamma \cup \overline{\Gamma}$. We adopt the following conventions: $\bar{\bar{a}} = a$ for all $a \in \Gamma$, $\bar{\varepsilon} = \varepsilon$ and for any word $u = \alpha_1 \cdots \alpha_n \in \widehat{\Gamma}^*$, $\bar{u} = \bar{\alpha_n} \cdots \bar{\alpha_1}$.

Let us consider the reduction system $S = \{(a\bar{a}, \varepsilon), (\bar{a}a, \varepsilon)\}_{a \in \widehat{\Gamma}}$. A word in $\widehat{\Gamma}^*$ is said to be **reduced** if it is S-reduced, i.e. it does not contain occurrences of $a\bar{a}$, $\bar{a}a$, for $a \in \Gamma$. As S is confluent, each word w is equivalent (mod $\leftrightarrow_S^*$) to a unique reduced word denoted $\rho(w)$. Note that for all $u \in \widehat{\Gamma}^*$, $\rho(u\bar{u}) = \rho(\bar{u}u) = \varepsilon$. Given a set X, we denote by $\rho(X)$ the set $\{\rho(x) \mid x \in X\}$.

The free group $\mathrm{F}(\Gamma)$ consists of reduced words over $\widehat{\Gamma}$. Its neutral element is the empty word and its product $\bullet$ is defined as $u \bullet v = \rho(uv)$.

The set of all words $u \in \widehat{\Gamma}^*$ such that $\rho(u) = \varepsilon$ is denoted $\mathcal{T}_\Gamma$. The Dyck language over Γ, denoted $\mathcal{D}_\Gamma$, is the set of all $u \in \mathcal{T}_\Gamma$, such that for every prefix $v \preccurlyeq u$: $\rho(v) \in \Gamma^*$. We will also write $\mathcal{D}_k$, $k \geq 1$, to refer to the set of Dyck words over any alphabet of size k.

2.2 ε-Reducible Context-Free Languages and Grammars

Definition 1. *An* **ε-Reducible Context-Free Grammar** *(ε-CFG) is a context free grammar $G = (N, T, S, P)$ (N is the set of nonterminal symbols, $\widehat{\Gamma}$ the terminal alphabet, $S \in N$ is the start symbol, and P is the set of productions) such that $T = \widehat{\Gamma}$ for some alphabet Γ and every production is in the form:*

$$X \longrightarrow \omega\Omega\bar{\omega}, \qquad \text{for } \omega \in \widehat{\Gamma}^*, \text{ and } \Omega \in N^*$$

For all $X \in N$, we define $\mathcal{L}_G(X) = \{u \in \widehat{\Gamma}^* \mid X \xrightarrow{*}_G u\}$; the language generated by G is $\mathcal{L}_G = \mathcal{L}_G(S)$.

An ε-Reducible Context-Free Language (ε-CFL) is a context-free language L that can be generated by an ε-CFG.

Example 2. Let $G = (N, \{\alpha, \beta, \bar{\alpha}, \bar{\beta}\}, S, P)$ be the ε-CFG whose productions are: $S \longrightarrow \beta X \bar{\beta}$, $X \longrightarrow \alpha X \bar{\alpha} + Y$, $Y \longrightarrow \bar{\alpha} Y Z \alpha + \bar{\beta}\beta$, $Z \longrightarrow \bar{\alpha} Z \alpha + \bar{\beta}\beta$.

One can easily check that:

$$\mathcal{L}_G(Z) = \bigcup_{n\geq 0} \bar{\alpha}^n \bar{\beta}\beta\alpha^n, \quad \mathcal{L}_G(Y) = \bigcup_{n\geq 0} \bar{\alpha}^n \beta\bar{\beta}(\Pi_{i=1}^n \mathcal{L}_G(Z)\alpha),$$
$$\mathcal{L}_G(S) = \beta\mathcal{L}_G(X)\bar{\beta}, \quad \mathcal{L}_G(X) = \bigcup_{n\geq 0} \alpha^n \mathcal{L}_G(Y)\bar{\alpha}^n.$$

It follows that: $\mathcal{L}_G = \bigcup_{n,m,r_1,\ldots r_m \geq 0} \beta\alpha^n\bar{\alpha}^m\bar{\beta}\beta(\Pi_{i=1}^m \bar{\alpha}^{r_i}\bar{\beta}\beta\alpha^{r_i+1})\bar{\alpha}^n\bar{\beta}$. □

It seems clear that every ε-CFL L satisfies $\rho(L) = \{\varepsilon\}$. One can indeed observe that every terminal word generated from a nonterminal symbol $X \in N$ reduce to ε. However, there are context-free languages that reduce to ε and which cannot be generated by an ε-CFG. We prove this by using a "pumping lemma" for ε-CFLs.

Lemma 3. *If $L \subseteq \widehat{\Gamma}^*$ is an ε-CFL, then there exists some integer $p \geq 1$ such that every word $s \in L$ with $|s| \geq p$ can be written as $s = uvwxy$ with*

1. *$\rho(uy) = \varepsilon$, $\rho(vx) = \varepsilon$ and $\rho(w) = \varepsilon$*
2. *$|vwx| \leq p$,*
3. *$|vx| \geq 1$, and*
4. *$uv^n wx^n y$ is in L for all $n \geq 0$.*

Proof (Sketch). Let G be an ε-CFG generated L. The proof of the pumping lemma for context-free languages is based on the fact that if a word $s \in L$ is long enough, there are a non terminal A and terminal words u, v, w, x, y such that $S \xrightarrow{*}_G uAy \xrightarrow{*}_G uvAxy \xrightarrow{*}_G uvwxy$ and $s = uvwxy$. Since G is an ε-CFG, this implies that $\rho(uy) = \varepsilon$, $\rho(vx) = \varepsilon$ and $\rho(w) = \varepsilon$ □

Proposition 4. *There is a context-free language L satisfying $\rho(L) = \varepsilon$ which is not an ε-CFL.*

Proof (Sketch). By applying Lemma 3 to the set L of words $(\alpha\bar{\alpha})^n\beta(\alpha\bar{\alpha})^n\bar{\beta}$ for $n \geq 0$, we can show that L is not an ε-CFL. □

Proposition 5. *The class of ε-CFLs is closed under union, intersection with regular sets, concatenation and Kleene star.*

Proof. Obviously, the class of ε-CFLs is closed under union, concatenation and Kleene star. Let us prove the closure under intersection with regular sets. Let L be generated by an ε-CFG $G = (N, \widehat{\Gamma}, P, S)$ and R be a regular language. There is a monoid morphism $\mu : \widehat{\Gamma}^* \to M$, where M is a finite monoid and $H \subseteq M$ such that $R = \mu^{-1}(H)$. We construct the ε-CFG $G' = (N', \widehat{\Gamma}, P', S')$ where $N' = \{X_m \mid X \in N, m \in M\} \cup \{S'\}$ and P' is the set of all productions:

- $X_m \longrightarrow \alpha X_{1,m_1} \cdots X_{n,m_n}\bar{\alpha}$ such that $X \longrightarrow \alpha X_1 \cdots X_n \bar{\alpha} \in P$ and $m = \mu(\alpha)m_1 \cdots m_n\mu(\bar{\alpha})$

– $S' \longrightarrow S_m$ for $m \in H$

Then for every $u \in \widehat{\Gamma}^*$, for every $X \in N$ and $m \in M$:

$$X_m \xrightarrow{*}_{G'} u \text{ iff } X \xrightarrow{*}_{G} u \text{ and } u \in \mu^{-1}(m).$$

It follows that $\mathcal{L}_{G'} = L \cap \mu^{-1}(H)$. □

2.3 A Chomsky–Schützenberger-like theorem for ε-CFLs

The Chomsky–Schützenberger theorem states that a language $L \subseteq \Sigma^*$ is context-free iff there is an alphabet Γ, a regular set $R \subseteq \widehat{\Gamma}^*$, and a homomorphism $h : \widehat{\Gamma}^* \to \Sigma^*$ such that

$$L = h(R \cap \mathcal{D}_\Gamma).$$

This implies that the whole class of context-free languages is generated by homomorphic images of ε-CFLs, since $R \cap \mathcal{D}_B$ is an ε-CFL. To get an homomorphic characterization for ε-CFLs, we introduce a class of homomorphisms under which the family of ε-CFLs is closed.

Definition 6. *A homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$ is said to be* symmetric *if for all $\alpha \in \widehat{\Sigma}$, $g(\bar{\alpha}) = \overline{g(\alpha)}$.*

Proposition 7. *The class of ε-CFLs is closed under symmetric homomorphism.*

Proof. Consider a language L generated by an ε-CFG $G = (N, \widehat{\Gamma}, P, S)$ and $g : \widehat{\Gamma}^* \to \widehat{\Sigma}^*$ be a symmetric homomorphism. We construct an ε-CFG $G' = (N, \widehat{\Sigma}, P', S)$ generating $g(L)$ as follows:

$P' = \{X \longrightarrow g(u)\Omega g(\bar{u}) \mid X \longrightarrow u\Omega\bar{u} \in P, \Omega \in N^*, u \in \widehat{\Gamma}^*\}$. □

More generally, ε-CFLs are closed under every homomorphism g satisfying "$\rho(u) = \varepsilon \implies \rho(g(u)) = \varepsilon$".
We can now state the main result of this section.

Theorem 8. *A set $L \subseteq \widehat{\Gamma}^*$ is an ε-CFL iff there is an alphabet Σ, a symmetric homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$, and a regular set $R \subseteq \widehat{\Sigma}^*$ such that*

$$L = g(R \cap \mathcal{D}_\Sigma).$$

The "if" part of Theorem 8 is direct using Propositions 5 and 7. The "only if" part is obtained using a slight adaptation of the proof of the non-erasing variant of the Chomsky–Schützenberger theorem given in [17].
We conclude this section by emphasizing that Theorem 8 and Propositions 5 and 7 provide another characterization of the class of ε-CFLs:

Corollary 9. *The family of ε-CFLs is the least family of languages that contains the Dyck language and is closed under union, intersection with regular sets, symmetric homomorphisms, concatenation and Kleene star.*

2.4 Related works

In [4], the authors define (pure) balanced grammars that are context-free grammars whose set of productions is a (possibly infinite) regular set of rules of the form $X \longrightarrow \alpha m \bar{\alpha}$, where $\alpha \in \Gamma$ and $m \in N^*$. Balanced grammars do not generate all ε-CFLs included in $\mathcal{D}_\Gamma$, for example they cannot generate the set $\{\beta(\alpha\bar{\alpha})^n(\gamma\bar{\gamma})^n\bar{\beta} \mid n \geq 0\}$.

Introduced in [15], input-driven languages, more recently known as Visibly Pushdown Languages (VPLs), are extensions of balanced languages defined over a structured alphabet: Σ_c is the set of call symbols, Σ_r the set of returns and Σ_ℓ the set of local symbols. They are recognized by pushdown automata that push onto the stack only when reading a call, pop the stack only on returns, and do not use the stack when reading local actions. The input word hence controls the permissible operations on the stack—however, there is no restriction on the symbols that can be pushed or popped. This implies that there are visibly pushdown languages which are not ε-CFLs. However the ε-CFL $\{(\alpha\bar{\alpha})^n(\beta\bar{\beta})^n \mid n \geq 0\}$ is not a VPL when $\Sigma_c = \Gamma$ and $\Sigma_r = \bar{\Gamma}$.

Also note that unlike ε-CFLs, VPLs are closed under intersection. We will see (Theorem 18) that the intersection of an ε-CFL with the Dyck language is an indexed language.

3 Epsilon-Reducible Context-Free Transductions

In this section, we extend the notion of ε-reducibility to transductions. We consider a subclass of context-free transductions such that their domains are ε-CFLs. We give a Nivat-like presentation of those transductions.

3.1 Transductions

We briefly introduce rational and context-free transductions. The reader can refer to [2] for a more detailed presentation.

Let Γ and Σ be two finite alphabets, we consider the monoid $\Gamma^* \times \Sigma^*$ whose product is the product on words, extended to pairs of words : $(u_1, v_1)(u_2, v_2) = (u_1u_2, v_1v_2)$. A subset τ of $\Gamma^* \times \Sigma^*$ is called a (Γ, Σ)-*transduction*.

Transductions are viewed as (partial) functions from Γ^* toward subsets of Σ^*: for any $u \in \Gamma^*$, $\tau(u) = \{v \in \Sigma^* \mid (u, v) \in \tau\}$. For every $L \subseteq \Gamma^*$, the *image* (or *transduction*) of L by τ is $\tau(L) = \bigcup_{u \in L} \tau(u)$. The *domain* of τ is $\mathrm{Dom}(\tau) = \{u \mid \exists v, (u, v) \in \tau\}$.

Rational transductions: A rational (Γ, Σ)-transduction is a rational subset of the monoid $\Gamma^* \times \Sigma^*$. Among the different characterizations of rational transductions, let us cite the Nivat theorem [16] stating that rational transductions are relations $\tau = \{(g(u), f(u)) \mid u \in R\}$, for some regular set R and homomorphisms f and g.

Rational transductions are closed by composition and many classes of languages are closed under rational transductions. In particular, $\tau(L)$ is rational if L is rational, and $\tau(L)$ is context-free if L is context-free.

Associated with the Nivat theorem, the Chomsky–Schützenberger theorem establish in a stronger version that a language L is context-free iff there is a rational transduction τ such that $L = \tau(\mathcal{D}_2)$.

Context-free transductions: Following [2, page 62], a transduction $\tau \subseteq \Gamma^* \times \Sigma^*$ is context-free if there is an alphabet A, a context-free language $K \subseteq A^*$ and two homomorphisms $f : A^* \to \Sigma^*$ and $g : A^* \to \Gamma^*$ such that $\tau = \{(g(u), f(u)) \mid u \in K\}$.

Equivalently, τ is context-free if it is generated by a context-free transduction grammar. This is a context-free grammar whose terminals are pairs of words. Derivations are done as usually but the product used on terminal pairs is the product of the monoid $\Gamma^* \times \Sigma^*$.

Context-free transductions enjoy however fewer good properties, in particular, [2, page 62] they are not closed under composition and classes of languages are usually not closed under them. For example, images of regulars languages are context-free languages and images of context-free languages are recursively enumerable languages.

3.2 ε-Reducible Context-Free Transductions and Transducers

Definition 10. *An* **ε-Reducible Context-Free Transduction Grammar** *(ε-CFTG) is a context-free transducer $G = (N, \widehat{\Gamma}, \Sigma, S, P)$ in which every production is in the form*

$$X \longrightarrow (\omega, u)\Omega(\bar{\omega}, v), \qquad \text{with } \omega \in \widehat{\Gamma}^*,\ u, v \in \Sigma^*,\ \Omega \in N^*.$$

The transduction generated by G is $\mathcal{T}_G = \{(u, v) \in \widehat{\Gamma}^* \times \Sigma^* \mid S \xrightarrow{*}_G (u, v)\}$. An ε-reducible context-free transduction (ε-CFT) is a context-free transduction generated by an ε-CFTG.

Example 11. Let $G = (N, \{\alpha, \beta, \bar{\alpha}, \bar{\beta}\}, \{a\}, S, P)$ be the ε-CFTG whose productions are:

$S \longrightarrow (\beta, \varepsilon)X(\bar{\beta}, \varepsilon) \quad X \longrightarrow (\alpha, \varepsilon)X(\bar{\alpha}, \varepsilon)\ X \longrightarrow (\varepsilon, \varepsilon)Y(\varepsilon, \varepsilon)$
$Y \longrightarrow (\bar{\alpha}, a)YZ(\alpha, \varepsilon)\ Z \longrightarrow (\bar{\alpha}, a)Z(\alpha, a)\ Y \longrightarrow (\bar{\beta}, \varepsilon)(\beta, \varepsilon)\ Z \longrightarrow (\bar{\beta}, \varepsilon)(\bar{\beta}, \varepsilon).$

Let τ be the transduction generated by G. The domain of τ is the ε-CFL given in Example 2 and one can easily check that

$$\tau = \bigcup_{n,m,r_1,\ldots r_m \geq 0} (\beta\alpha^n\bar{\alpha}^m\bar{\beta}\beta(\Pi_{i=1}^m \bar{\alpha}^{r_i}\bar{\beta}\beta\alpha^{r_i+1})\bar{\alpha}^n\bar{\beta}, a^{m+2r_1+\cdots+2r_m}).$$

□

Theorem 12. *Given a transduction $\tau \subseteq \widehat{\Gamma}^* \times A^*$, the following properties are equivalent:*

1. *τ is an ε-reducible context-free transduction;*

2. *there is an alphabet* Δ*, an* ε*-CFL* $X \subseteq \widehat{\Delta}^*$*, a symmetric homomorphism* $g : \widehat{\Delta}^* \to \widehat{\Gamma}^*$ *and a homomorphism* $h : \widehat{\Delta}^* \to A^*$ *such that*

$$\tau = \{(g(u), h(u)) \mid u \in X\};$$

3. *there is an alphabet* Δ*, a symmetric homomorphism* $g : \widehat{\Delta}^* \to \widehat{\Gamma}^*$*, a homomorphism* $h : \widehat{\Delta}^* \to A^*$ *and a regular set* $R \subseteq \widehat{\Delta}^*$ *such that*

$$\tau = \{(g(u), h(u)) \mid u \in R \cap \mathcal{D}_\Delta\}.$$

Proof. $(1 \Rightarrow 2)$ Suppose τ to be generated by an ε-CFTG $G = (N, \widehat{\Gamma}, \Sigma, S, P)$. We define the ε-CFG $G' = (N, \widehat{\Delta}, S, P')$ where $\Delta = P$, and the set of productions of P' is obtained by transforming every $p : X \longrightarrow (\omega, v)\Omega(\bar{\omega}, w) \in P$ into $X \longrightarrow p\Omega\bar{p}$. Now, let $h : \widehat{\Delta}^* \to A^*$ and $g : \widehat{\Delta}^* \to \widehat{\Gamma}^*$ such that for every $p : X \longrightarrow (\omega, v)\Omega(\bar{\omega}, w) \in P$, $g(p) = \omega$, $g(\bar{p}) = \bar{\omega}$ and $h(p) = v$, $h(\bar{p}) = w$. Clearly we have

$$\mathcal{T}_G = \{(g(u), h(u)) \mid u \in \mathcal{L}(G')\}.$$

$(2 \Rightarrow 3)$ Suppose that $\tau = \{(g(u), h(u)) \mid u \in X\}$ where X is an ε-CFL and g symmetric. From Theorem 8, there is an alphabet C, a regular set $R \subseteq \widehat{C}^*$, and a symmetric homomorphism $g' : \widehat{C}^* \to \widehat{\Delta}^*$ such that $X = g'(R \cap \mathcal{D}_\Delta)$. The homomorphism $g \circ g'$ is symmetric as g and g' are both symmetric and $\tau = \{(g(g'(x)), h(g'(x))) \mid x \in R \cap \mathcal{D}_\Delta\}$.

$(3 \Rightarrow 1)$ Let $\tau = \{(g(u), h(u)) \mid u \in R \cap \mathcal{D}_\Delta\}$ where R is a regular language and g is symmetric. From Proposition 5, $R \cap \mathcal{D}_\Delta$ is an ε-CFL. Let us suppose that $R \cap \mathcal{D}_\Delta$ is generated by the ε-CFG $G = (N, \widehat{\Delta}, P, S)$, then τ is generated by the ε-CFTG $G' = (N, \widehat{\Gamma}, \widehat{\Sigma}P', S)$ where

$$P' = \{X \longrightarrow (g(u), f(u))\Omega(g(\bar{u}), h(\bar{u})) \mid X \longrightarrow u\Omega\bar{u} \in P, \Omega \in N^*, u \in \widehat{\Gamma}^*\}.$$

□

Theorem 12 implies that the image of a set X by an ε-CFT can be represented as $h(g^{-1}(X) \cap R \cap \mathcal{D}_\Delta)$ with R being a regular set, h a morphism and g a symmetric morphism. It is then clear that the family of images of regular sets by ε-CFTs is the family of context-free languages; we will see (Theorem 15) that the family of images of the Dyck language is that of indexed languages, but more generally, images of ε-CFLs by ε-CFTs are recursively enumerable languages.

Proposition 13. *Given a recursively enumerable language* E*, there is an* ε*-CFT* τ*, and an* ε*-CFL* Z *such that* $E = \tau(Z)$*.*

Proof. Let $E \subseteq \Sigma^*$. From [10], there is an alphabet Γ, a homomorphism $h : \widehat{\Gamma}^* \to \Sigma^*$, and a context-free language $K \subseteq \widehat{\Gamma}^*$ such that $E = h(K \cap \mathcal{D}_\Gamma)$.

Let $g : \widehat{\Gamma}^* \to \widehat{\Gamma}^*$ be the injective symmetric homomorphism defined by $x \mapsto x\bar{x}$, for all $x \in \widehat{\Gamma}$. Then $E = h(g^{-1}(Z) \cap \mathcal{D}_\Gamma)$, for $Z = g(K)$, that is, from Theorem 12, $E = \tau(Z)$, where τ is an ε-CFT. Note finally that Z is an ε-CFL: from the grammar in Chomsky normal form generating K, one obtain an ε-CFG generating Z by replacing the terminal productions $X \longrightarrow a$ by $X \longrightarrow g(a)$. □

4 Characterizations of Indexed Languages

In this final section, we relate indexed languages to ε-CFTs by showing that indexed language are sets $\tau(\mathcal{D}_2)$, where τ is an ε-CFT. This gives rise to various homomorphic characterizations of indexed languages.

4.1 Indexed Grammars and Languages

Introduced by Aho[1], indexed grammars extend context-free grammars by allowing nonterminals to yield a stack. Derivable elements are then represented by symbols X^ω where X is a nonterminal and ω is a word called *index word*. Index words are accessed by a FIFO process: during a step of derivation of X^ω, it is possible to add a symbol in head ω, or to remove its first letter. Additionally, ω can be duplicated and distributed over other nonterminals.

Formally, an **indexed grammar** is a structure $\mathcal{I} = (N, I, \Sigma, S, P)$, where N is the set of nonterminals, Σ is the set of terminals, $S \in N$ is the start symbol, I is a finite set of indexes, and P is a finite set of productions of the form

$$X_0{}^{\eta_0} \longrightarrow u_0 X_1{}^{\eta_1} u_1 \cdots X_n{}^{\eta_n} u_n$$
$$\text{with } u_i \in \Sigma^*,\ X_i \in N \text{ and } \eta_i \in I \cup \{\varepsilon\} \text{ for } i \in \{0, \ldots, n\}.$$

Indexes are denoted as *superscript*, and we do not write indexes equal to ε.

Sentences are words $u_1 A_1{}^{\omega_1} \ldots u_n A_n{}^{\omega_n} u_{n+1}$ with $u_i \in \Sigma^*$, $A_i \in N$ and $\omega_i \in I^*$. The derivation rule "$\longrightarrow_{\mathcal{I}}$" is a binary relation over sentences defined by

$$\Omega_1 A^{\eta\omega} \Omega_2 \longrightarrow_{\mathcal{I}} \Omega_1 u_0 B_1{}^{\eta_1\omega} \cdots B_n{}^{\eta_1\omega} u_n \Omega_2$$
$$\text{iff there is a production } A^\eta \longrightarrow u_0 B_1{}^{\eta_1} u_1 \ldots B_n{}^{\eta_n} u_n \in P.$$

The language generated by $\mathcal{I}$ is $\mathcal{L}_{\mathcal{I}} = \{u \in \Sigma^* \mid S \xrightarrow{*}_{\mathcal{I}} u\}$. Languages generated by indexed grammars are called **indexed languages**.

Example 14. Let us consider the following indexed grammar $\mathcal{I} = (N, I, A, S, P)$ with $N = \{S, X, A, B, C\}$, $I = \{\beta, \alpha\}$, $A = \{a, b, c\}$ and P consists of the following rules:

$$\begin{array}{llll} p_1 : S \longrightarrow X^\beta, & p_2 : S \longrightarrow \varepsilon, & p_3 : X \longrightarrow X^\alpha, & p_4 : X \longrightarrow ABC, \\ p_5 : A^\alpha \longrightarrow aA, & p_6 : A^\beta \longrightarrow \varepsilon, & p_7 : B^\alpha \longrightarrow bB, & p_8 : B^\beta \longrightarrow \varepsilon, \\ p_9 : C^\alpha \longrightarrow cC, & p_{10} : C^\beta \longrightarrow \varepsilon. & & \end{array}$$

Here is a possible derivation:

$$S \xrightarrow{p_1}_{\mathcal{I}} X^\beta \xrightarrow{p_3}_{\mathcal{I}} X^{\alpha\beta} \xrightarrow{p_3}_{\mathcal{I}} X^{\alpha\alpha\beta} \xrightarrow{p_4}_{\mathcal{I}} A^{\alpha\alpha\beta} B^{\alpha\alpha\beta} C^{\alpha\alpha\beta} \xrightarrow{p_5}_{\mathcal{I}} aA^{\alpha\beta} B^{\alpha\alpha\beta} C^{\alpha\alpha\beta}$$
$$\xrightarrow{p_5}_{\mathcal{I}} aaA^\beta B^{\alpha\alpha\beta} C^{\alpha\alpha\beta} \xrightarrow{p_6}_{\mathcal{I}} aaB^{\alpha\alpha\beta} C^{\alpha\alpha\beta} \xrightarrow{p_7 p_7 p_8}_{\mathcal{I}} aabbC^{\alpha\alpha\beta} \xrightarrow{p_9 p_9 p_{10}}_{\mathcal{I}} aabbcc$$

The language generated by $\mathcal{I}$ is $\{a^n b^n c^n, n \geq 0\}$. □

4.2 Characterizations of Indexed Languages

We provide now homomorphic characterizations of indexed languages by establishing a strong connexion between indexed languages and ε-CFTs.

Theorem 15. *A language L is indexed iff there is an ε-CFT τ such that*

$$L = \tau(\mathcal{D}_2).$$

Let us informally explain the proof of Theorem 15. First we need to consider normal forms of indexed grammars (which extend the normal form given in [1]) and ε-CFT grammars.

An indexed grammar is said to be *reduced* if its productions are in the forms:

$$X_0 \longrightarrow uX_1{}^{\alpha} \cdots X_n{}^{\alpha} v, \qquad \text{or } X_0{}^{\alpha} \longrightarrow uX_1 \cdots X_n v;$$
$$\text{with } n \geq 0,\ X_i \in N,\ u, v \in \Sigma^* \text{ and } \alpha \in I \cup \{\varepsilon\}.$$

An ε-CFTG is said to be *reduced* if its productions are in the form:

$$X_0 \longrightarrow (\alpha, u)\Omega(\bar{\alpha}, v) \text{ with } \Omega \in N^*,\ u, v \in \Sigma^* \text{ and } \alpha \in \widehat{\Gamma} \cup \{\varepsilon\}.$$

Let us consider the bijective mapping φ that maps a reduced indexed grammar $\mathcal{I} = (N, I, \Sigma, P, S)$ into a reduced ε-CFTG $\varphi(\mathcal{I}) = (N, \widehat{I}, \Sigma, \varphi(P), S)$ by transforming every production

$$p : X_0 \longrightarrow uX_1{}^{\alpha} \cdots X_n{}^{\alpha} v \text{ into } \varphi(p) : X_0 \longrightarrow (\alpha, u)X_1 \cdots X_n(\bar{\alpha}, v), \text{ and}$$
$$p : X_0{}^{\alpha} \longrightarrow uX_1 \cdots X_n v \quad \text{into } \varphi(p) : X_0 \longrightarrow (\bar{\alpha}, u)X_1 \cdots X_n(\alpha, v).$$

The idea behind the construction is to write, into the terminal inputs of the ε-CFTG, the index operations made by the indexed grammar. The transduction grammar thus created is able to capture every index modifications of the initial indexed grammar, but also accepts bad computations. We claim that by restricting the domain to Dyck words, we exactly get derivations equivalent to those of the indexed grammar.

For example, there would be a derivation

$$X \longrightarrow u_1 X_1{}^{\alpha} v_1 \longrightarrow u_2 Y_1{}^{\alpha} Y_2{}^{\alpha} v_2 \longrightarrow u_3 Y_1{}^{\alpha} w_3 Z v_3$$

in $\mathcal{I}$ iff there was a derivation of the following form in $\varphi(\mathcal{I})$:

$$X \longrightarrow (\alpha, u_1)X_1(\bar{\alpha}, v_1) \longrightarrow (\alpha, u_2)Y_1Y_2(\bar{\alpha}, v_2) \longrightarrow (\alpha, u_3)Y_1(\bar{\alpha}, w_3)Z(\alpha\bar{\alpha}, v_3).$$

Claim: *There is a derivation $S \overset{*}{\longrightarrow}_{\mathcal{I}} v_1 Y_1{}^{w_1} v_2 \cdots Y_n{}^{w_n} v_{n+1}$ iff there is a derivation $S \overset{*}{\longrightarrow}_{\varphi(\mathcal{I})} (u_1, v_1)Y_1(u_2, v_2) \cdots Y_n(u_{n+1}, v_{n+1})$ where $u_1 \cdots u_{n+1}$ belongs to $\mathcal{D}_I$ and $\rho(u_1 \cdots u_i) = w_i^R$ for $i \in \{1, \ldots, n\}$ (w_i^R is the mirror image of w_i).*

This can be proved by induction over the length of derivations, and implies that $\mathcal{T}_{\varphi(\mathcal{I})}(\mathcal{D}_I) = \mathcal{L}_{\mathcal{I}}$. Because of the bijectivity of the construction, we obtain:

"A language L is indexed iff there is an ε-CFT τ and $k \in \mathbb{N}$ s.t. $L = \tau(\mathcal{D}_k)$."

Finally, it is possible to define from every ε-CFT τ, an ε-CFT τ' such that $\tau(\mathcal{D}_\Gamma) = \tau'(\mathcal{D}_2)$, by encoding every $\alpha_i \in \Gamma$ by a word 01^i0 and $\bar{\alpha}_i$ by $\bar{0}\bar{1}^i\bar{0}$.

Example 16. Let $\mathfrak{I} = (N, I, \Sigma, S, P)$ be an indexed grammar with $N = \{S, X, Y, W, Z\}$, $I = \{\beta, \alpha\}$, $A = \{a\}$ and P consists of the rules:

$S \longrightarrow X^{\beta}$, $X \longrightarrow X^{\alpha}$, $X \longrightarrow Y$, $Y^{\alpha} \longrightarrow aYZ$

$Y^{\beta} \longrightarrow \varepsilon$, $Z^{\alpha} \longrightarrow aZa$, $Z^{\beta} \longrightarrow \varepsilon$.

Initially defined in [7], the grammar $\mathfrak{I}$ generates the language $L = \{a^{n^2} \mid n \geq 0\}$.

Applying the bijection φ defined above to $\mathfrak{I}$, we get the ε-CFTG G given in Example 11 and generating the transduction

$$\tau = \bigcup_{n,m,r_1,\ldots r_m \geq 0} (\beta\alpha^n\bar{\alpha}^m\bar{\beta}\beta(\Pi_{i=1}^m \bar{\alpha}^{r_i}\bar{\beta}\beta\alpha^{r_i+1})\bar{\alpha}^n\bar{\beta}, a^{m+2r_1+\cdots+2r_m}).$$

For every $u = \beta\alpha^n\bar{\alpha}^m\bar{\beta}\beta(\Pi_{i=1}^m \bar{\alpha}^{r_i}\bar{\beta}\beta\alpha^{r_i+1})\bar{\alpha}^n\bar{\beta} \in \mathrm{Dom}(\tau)$,

$$\begin{aligned} u \text{ is a Dyck word} &\implies m = n,\ r_1 = 0, \text{ and for all } i \in [0, m-1],\ r_{i+1} = r_i + 1 \\ &\implies \tau(u) = a^{n+2(0+1+\cdots+n-1)} \\ &\implies \tau(u) = a^{n^2}. \end{aligned}$$

It follows that $\tau(\mathcal{D}_I) = \{a^{n^2}\}_{n\geq 0} = \mathcal{L}_{\mathfrak{I}}$. □

Corollary 17. *A language L is indexed if there is a homomorphism h, a symmetric homomorphism g, a regular set R and $k \in \mathbb{N}$ such that*

$$L = h(g^{-1}(\mathcal{D}_2) \cap R \cap \mathcal{D}_k).$$

Theorem 18. *A language L is indexed iff there is an ε-CFL K, a morphism h, and an alphabet Γ such that*

$$L = h(K \cap \mathcal{D}_\Gamma).$$

Proof. ($\Rightarrow$) Let $L \subseteq A^*$ be an indexed language. From Theorem 15 and Theorem 12, there are alphabets Σ, Γ, an ε-CFL $K \subseteq \widehat{\Gamma}^*$, a homomorphism $h : \widehat{\Gamma^*} \to A^*$ and a symmetric homomorphism $g : \widehat{\Gamma}^* \to \widehat{\Sigma}^*$ such that $L = h(K \cap g^{-1}(\mathcal{D}_\Sigma))$. We suppose that $\Sigma \cap A = \emptyset$ (otherwise, it suffices to work with a copy of Σ), and define the homomorphism $\mu : \widehat{\Gamma}^* \to \widehat{\Delta}^*$, for $\Delta = \Sigma \cup A$, by $\alpha \mapsto g(\alpha)h(\alpha)\overline{h(\alpha)}$. For all $u \in \widehat{\Gamma}^*$, $\mu(u) \in \mathcal{D}_\Delta$ iff $u \in g^{-1}(\mathcal{D}_\Sigma)$; in addition, $\pi_A(\mu(u)) = h(u)$, with π_A being the projection of $\widehat{\Delta}^*$ into A^*. Then we have:

$$\pi_A(\mu(K) \cap \mathcal{D}_\Delta) = h(K \cap g^{-1}(\mathcal{D}_\Sigma)) = L.$$

Now, as μ satisfies "$\rho(u) = \varepsilon \implies \rho(g(u)) = \varepsilon$" and K is an ε-CFL, so is $\mu(K)$.

($\Leftarrow$) Obvious from Theorem 15 and Proposition 12, by choosing g to be the identity mapping. □

Acknowledgement

The authors would like to thank Pr. Jean-Marc Talbot whose remarks and suggestions greatly improved the development of this paper.

References

1. Aho, A.: Indexed grammars–an extension of context-free grammars. J. ACM 15, 647–671 (1968)
2. Berstel, J.: Transductions and Context-Free Languages. Teubner Verlag (1979)
3. Berstel, J., Boasson, L.: In: Handbook of Theoretical Computer Science (Vol. B), chap. Context-free Languages, pp. 59–102. MIT Press (1990)
4. Berstel, J., Boasson, L.: Balanced grammars and their languages. In: Formal and Natural Computing. Lecture Notes in Comput. Sci., vol. 2300, pp. 3–25. Springer (2002)
5. Caucal, D.: On infinite transition graphs having a decidable monadic theory. Theor. Comput. Sci. 290(1), 79–115 (2003)
6. Chomsky, N., Schützenberger, M.P.: The algebraic theory of context-free languages. In: Computer Programming and Formal Systems, pp. 118–161. Studies in Logic, North-Holland Publishing (1963)
7. Fischer, M.J.: Grammars with macro-like productions. In: 9th Annual Symposium on Switching and Automata Theory. pp. 131–142. IEEE Computer Society (1968)
8. Fratani, S.: Automates à piles de piles ... de piles. Ph.D. thesis, Université Bordeaux 1 (2005)
9. Hague, M., Murawski, A.S., Ong, C.L., Serre, O.: Collapsible pushdown automata and recursion schemes. In: LICS, Proceedings. pp. 452–461. IEEE Computer Society (2008)
10. Hirose, S., Nasu, M.: Left universal context-free grammars and homomorphic characterizations of languages. Information and Control 50(2), 110 – 118 (1981)
11. Kanazawa, M.: Multidimensional trees and a Chomsky–Schützenberger–Weir representation theorem for simple context-free tree grammars. J. of Logic and Computation (2014)
12. Knapik, T., Niwinski, D., Urzyczyn, P.: Higher-order pushdown trees are easy. In: FOSSACS. Lecture Notes in Comput. Sci., vol. 2303, pp. 205–222. Springer (2002)
13. Maslov, A.N.: Hierarchy of indexed languages of arbitrary level. Soviet Math. Dokl 115(14), 1170–1174 (1974)
14. Maslov, A.N.: Multilevel stack automata. Problems of Information Transmission 12, 38–43 (1976)
15. Mehlhorn, K.: Pebbling mountain ranges and its application to DCFL-recognition. Automata, Languages and Programming 85, 422–435 (1980)
16. Nivat, M.: Transductions des langages de Chomsky. Ann. Inst. Fourier 18(1), 339–455 (1968)
17. Okhotin, A.: Non-erasing variants of the chomsky-schützenberger theorem. In: Developments in Language Theory. Lecture Notes in Comput. Sci., vol. 7410, pp. 121–129. Springer (2012)
18. Ong, L.: Higher-order model checking: An overview. In: LICS. pp. 1–15. IEEE Computer Society (2015)
19. Sorokin, A.: Monoid automata for displacement context-free languages. CoRR abs/1403.6060 (2014)
20. Weir, D.: Characterizing Mildly Context-Sensitive Grammar Formalisms. Ph.D. thesis, University of Pennsylvania (1988), available as Technical Report MS-CIS-88-74