Contents lists available at ScienceDirect

Information and Computation

www.elsevier.com/locate/yinco

Epsilon-reducible context-free languages and characterizations of indexed languages

Séverine Fratani *, El Makki Voundy

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

ARTICLE INFO

Article history:
Received 2 July 2016
Received in revised form 19 September 2017
Accepted 22 July 2018
Available online xxxx

Keywords:
Context-free grammars
Homomorphic characterizations
Transductions
Indexed languages

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.

1. Introduction

The well known Chomsky–Schützenberger theorem [1] 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. Equivalently, every context-free language can be written $L = h(g^{-1}(\mathcal{D}_2) \cap R)$, for some regular set R, and homomorphisms h and g. Combined with Nivat's characterization of rational transductions, this means that any context-free language can be obtained by applying a rational transduction to $\mathcal{D}_2$.

Let us consider wider families of languages. Maslov defines in [2] an infinite hierarchy of languages included in recursively enumerable languages. The level 1 consists of context-free languages and the level 2 of indexed languages (initially defined by Aho [3]). Known as higher order languages since the last decades, the languages of the hierarchy and derived objects as higher order trees [4], higher order schemes [5], or higher order graphs [6], are used to model programming languages and are in the core of recent researches in program verification [7].

It is stated in [8] and proved in [9] that each level $\mathcal{L}_k$ of the hierarchy is a principal full trio 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

* Corresponding author.
E-mail addresses: severine.fratani@lif.univ-mrs.fr (S. Fratani), makki.voundy@lif.univ-mrs.fr (E.M. Voundy).

https://doi.org/10.1016/j.ic.2019.104444
0890-5401/© 2019 Elsevier Inc. All rights reserved.

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). Informally, there are context-free languages defined over parenthesis alphabets ($\Gamma \cup \bar{\Gamma}$ where $\bar{\Gamma}$ is a copy of Γ) and generated by context-free grammars whose productions have the form

$$X \longrightarrow \alpha X_1 \dots X_n \bar{\alpha} \text{ or } X \longrightarrow \bar{\alpha} X_1 \dots X_n \alpha,\ \alpha \in \Gamma,\ n \geq 0.$$

Then every ε-CFL is included in the set $\mathcal{T}_\Gamma$ of two-sided Dyck words over Γ: words that reduce to ε by the reduction $\{a\bar{a} \to \varepsilon, \bar{a}a \to \varepsilon\}_{a \in \Gamma}$. Note that ε-CFLs are a generalization of pure balanced context-free languages defined in [11], which are languages generated by grammars whose set of productions is a (possibly infinite) regular set of rules of the form $X \longrightarrow \alpha X_1 \dots X_n \bar{\alpha}$, where $\alpha \in \Gamma$ and $n \geq 1$.

We extend ε-CFLs 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 generalizations of the Chomsky–Schützenberger theorem. Indexed languages are:

- the images of $\mathcal{D}_2$ by ε-reducible context-free transductions (Theorem 5.6);
- sets $h(Z \cap \mathcal{D}_k)$; where k is an integer, Z an ε-CFL, and h a homomorphism (Theorem 5.10).

Beyond these main results, we study the class of ε-CFLs: we explore closure properties and express ε-CFLs using *ε-safe homomorphisms* which are homomorphisms that preserve the class. We establish a Chomsky–Schützenberger-like Theorem: ε-CFLs are languages that can be represented as $L = g(R \cap \mathcal{D}_k)$ for some integer k, regular language R, and ε-safe homomorphism g. We also prove that there exist context-free languages included in $\mathcal{T}_\Gamma$ that are not ε-CFLs, and establish some undecidability results.

Concerning ε-CFTs, we give a Nivat-like characterization: there 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 ε-safe homomorphism g. This leads to another homomorphic characterization of indexed languages: there are languages that can be written $h(g^{-1}(\mathcal{D}_2) \cap R \cap \mathcal{D}_k)$, for some integer k, regular language R, homomorphism h and ε-safe homomorphism g (Corollary 5.9).

Related works. Similar homomorphic characterizations have been given for subclasses of indexed languages: by Weir [12] for linear indexed languages, by Kanazawa [13] and Sorokin [14] for yields of tree languages generated by simple context-free grammars. The main difference is that in their cases, the homomorphism g is not ε-safe, but is *fixed* in function of k.

Homomorphic characterizations of indexed languages can be found in [15–17]. In particular, in [15], authors introduced the class of ε-CFLs by means of grammars that can be viewed as normal forms of the grammars presented here, and proposed homomorphic characterizations similar to ours. However, we propose here a broader and more general approach.

In [18], authors prove that the family of linear indexed languages form a principal full trio and give a generator of the family.

Overview. Section 1 introduces a few notations. Section 2 is devoted to the study of ε-CFLs. After introducing necessary notions, we define the class of ε-CFLs by means of context-free grammars. We then study their closure properties, and give a Chomsky–Schützenberger-like characterization. We then compare ε-CFLs with the class of two-sided Dyck context-free languages (TCFLs), which are context-free languages included in $\mathcal{T}_\Gamma$. We conclude by exploring some decidability problems. In Section 3, we extend our definition to transductions: we define the class of ε-CFTs which are context-free transductions whose domain is an ε-CFL. After a subsection giving background on transductions, we give a Nivat-like characterization of ε-CFTs. Section 4 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 several homomorphic characterizations. We also define an indexed language that generate the full trio of indexed languages. In the last section, we define the class of two-sided Dyck context-free transducers (TCFT) which are context-free transductions whose domain is a TCFL. We prove that a language is indexed iff it is the image of the Dyck language by a TCFT.

2. Notations

Throughout the paper, we will use the following conventions. Every mapping $f : A \to B$, will by extend into a map over subsets of A: for all $X \subseteq A$, $f(A) = \{f(x) \mid x \in X\}$.

Given two alphabets A and B, where $B \subseteq A$, $\pi_{A,B} : A^* \to B^*$ is the morphism defined by $a \notin B \mapsto \varepsilon$, and $a \in B \mapsto a$. By abuse of notation, we will often use the notation π_B, rather than $\pi_{A,B}$, without specify the domain alphabet A.

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

3. Epsilon-reducible context-free languages

We introduce the class of ε-reducible Context-Free Languages (ε-CFLs). Each ε-CFL over an alphabet Γ is a language over the alphabet $\Gamma \cup \overline{\Gamma}$ where $\overline{\Gamma}$ is a copy of Γ, and is included in the set of two-sided Dyck words over Γ, that is, in the set of words that reduce to ε via the rewriting system $\{\alpha\bar{\alpha} \longrightarrow \varepsilon,\ \bar{\alpha}\alpha \longrightarrow \varepsilon \mid \alpha \in \Gamma\}$. We propose a rather complete study of this class: characterizations, closure properties, a Chomsky–Schützenberger-like homomorphic characterization, comparison with other classes and decidability problems. We assume the reader to be familiar with context-free grammars and languages (see [19] for example), and present below a few necessary notions on Dyck languages.

3.1. Dyck and two-sided 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 systems $S = \{a\bar{a} \to \varepsilon, \bar{a}a \to \varepsilon \mid a \in \Gamma\}$ and $S_+ = \{a\bar{a} \to \varepsilon \mid a \in \Gamma\}$. As S and S_+ are confluent, each word w is equivalent (mod $\leftrightarrow_S^*$) to a unique S-irreducible word denoted $\rho(w)$ and is equivalent (mod $\leftrightarrow S_+{}^*$) to a unique S_+-irreducible word denoted $\rho^+(w)$.

For instance, if $w = a\bar{b}c\bar{c}\bar{d}db$, then $\rho(w) = a$ and $\rho^+(w) = a\bar{b}b$. Note that for all $u \in \widehat{\Gamma}^*$, $\rho(u\bar{u}) = \rho(\bar{u}u) = \varepsilon$. The confluence of the systems S and S_+ implies that:

Lemma 3.1. *For all $u_1, u_2, u_3 \in \widehat{\Gamma}^*$, $\rho(u_1u_2u_3) = \rho(u_1\rho(u_2)u_3)$ and $\rho^+(u_1u_2u_3) = \rho^+(u_1\rho^+(u_2)u_3)$.*

The set of all words $u \in \widehat{\Gamma}^*$ such that $\rho(u) = \varepsilon$ is denoted $\mathcal{T}_\Gamma$; it is the so-called two-sided Dyck language over Γ. 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 of u: $\rho(v) \in \Gamma^*$, or equivalently, the set of all $u \in \mathcal{T}_\Gamma$, such that $\rho^+(u) = \varepsilon$. We will also write $\mathcal{D}_k$, to refer to the set of Dyck words over any alphabet of size $k \geq 1$.

The *decomposition properties of Dyck words and two-sided Dyck words* gives an inductive definition of the languages $\mathcal{D}_\Gamma$ and $\mathcal{T}_\Gamma$ that will allow to make proofs by structural induction.

Lemma 3.2 *(Decomposition properties). Let $u \in \widehat{\Gamma}^*$, then $u \in \mathcal{D}_\Gamma$ (resp. $u \in \mathcal{T}_\Gamma$) iff one of the following case holds:*

1. *$u = \varepsilon$;*
2. *there are two nonempty words $u_1, u_2 \in \mathcal{D}_\Gamma$ (resp. $u_1, u_2 \in \mathcal{T}_\Gamma$) such that $u = u_1u_2$;*
3. *there is a symbol $\alpha \in \Gamma$ (resp. $\alpha \in \widehat{\Gamma}$) and a word $v \in \mathcal{D}_\Gamma$ (resp. $v \in \mathcal{T}_\Gamma$) such that $u = \alpha v\bar{\alpha}$.*

Proof. We prove the case of two-sided Dyck words, that of Dyck words is similar. Because of Lemma 3.1, every word $u \in \widehat{\Gamma}^*$ satisfying one of the cases 1, 2 or 3 belongs to $\mathcal{T}_\Gamma$. Conversely, let us consider a word $u \in \mathcal{T}_\Gamma$. If $u = \varepsilon$ then u satisfies the case 1. Otherwise, $u = \alpha v$, with $\alpha \in \widehat{\Gamma}$ and $v \in \widehat{\Gamma}^*$. From definition, there exists then a decomposition of v in $v = v_1\bar{\alpha}v_2$ with $v_1, v_2 \in \mathcal{T}_\Gamma$. In particular, if $v_2 \neq \varepsilon$, then u satisfies the case 2 (by choosing u_1 to be $\alpha v_1\bar{\alpha}$ and u_2 to be v_2). Otherwise, it satisfies the case 3. □

3.2. Epsilon-reducible context-free languages and grammars

We introduce the class of ε-reducible context-free languages using a syntactic restriction of context-free grammars. Let us first fix a few notations regarding context-free grammars.

A context-free grammar (CFG) is a structure $G = (N, T, S, P)$ where N is the set of non-terminals, T the set of terminals, S the initial non-terminal and P the set of productions. We call *sentence* a word over $N \cup T$ (denoted by symbols $\Omega, \Omega_1, \Omega_2, \ldots$). The set of terminal words derived from a non-terminal $X \in N$ is

$$\mathcal{L}_G(X) = \{u \in T^* \mid X \xrightarrow{*}_G u\}.$$

The context-free language (CFL) generated by G is then $\mathcal{L}_G = \mathcal{L}_G(S)$.

Recall that every CFL can be generated by a context-free grammar in (weak) Chomsky normal form. That is, such that every production $X \longrightarrow \Omega$ satisfies $\Omega \in N^* \cup \Sigma^*$.

Definition 3.3. An **ε-Reducible Context-Free Grammar** (ε-CFG) is a context free grammar $G = (N, T, S, P)$ such that $T = \widehat{\Gamma}$ for some alphabet Γ and every production is of the form:

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

An ε-Reducible Context-Free Language (ε-CFL) is a language that can be generated by an ε-CFG. The class of all ε-CFLs included in $\widehat{\Gamma}^*$ is denoted $ECFL(\Gamma)$.

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

The sets $\mathcal{D}_\Gamma$ and $\mathcal{T}_\Gamma$ belong to *ECFL*(Γ). Indeed, using the Decomposition property, one can easily checked that $\mathcal{D}_\Gamma$ is generated by the productions

$$S \longrightarrow SS \mid aS\bar{a} \mid \varepsilon, \text{ for } a \in \Gamma,$$

and that $\mathcal{T}_\Gamma$ is generated by the productions

$$S \longrightarrow SS \mid \alpha S\bar{\alpha} \mid \varepsilon, \text{ for } \alpha \in \widehat{\Gamma}.$$

The following will be used as running example throughout the paper.

Example 3.4. Let $G = (N, \{\alpha, \beta, \bar{\alpha}, \bar{\beta}\}, S, P)$ be the ε-CFG whose productions are:

$$S \longrightarrow \beta X\bar{\beta}, \qquad X \longrightarrow \alpha X\bar{\alpha} + Y, \qquad Y \longrightarrow \bar{\alpha} Y Z\alpha + \bar{\beta}\beta, \qquad Z \longrightarrow \bar{\alpha} Z\alpha + \bar{\beta}\beta.$$

One can easily check that:

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

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

We establish now that ε-CFLs can be characterized using a semantic restriction of context-free grammars: a language is an ε-CFL iff it can be generated by a grammar G such that for every non-terminal X, $\mathcal{L}_G(X) \subseteq \mathcal{T}_\Gamma$. This implies in particular that every ε-CFL is included in $\mathcal{T}_\Gamma$.

Proposition 3.5. *Given $L \subseteq \widehat{\Gamma}^*$, the following properties are equivalent:*

1. *L is an ε-CFL;*
2. *there is a context-free grammar $G = (N, \widehat{\Gamma}, P, S)$ such that $\mathcal{L}_G = L$ and for all $X \in N$: $\mathcal{L}_G(X) \subseteq \mathcal{T}_\Gamma$;*
3. *there is a context-free grammar $G = (N, \widehat{\Gamma}, P, S)$ such that $\mathcal{L}_G = L$ and for all $X \longrightarrow \Omega \in P$: $\pi_{\widehat{\Gamma}}(\Omega) \in \mathcal{T}_\Gamma$.*

Proof. Assertion $(1 \Rightarrow 2)$ comes easily from Lemma 3.1.

For $(2 \Rightarrow 3)$ we consider a context-free grammar $G = (N, \widehat{\Gamma}, P, S)$ such that for all $X \in N$: $\mathcal{L}_G(X) \subseteq \mathcal{T}_\Gamma$. Removing unused productions and nonterminals, we can also suppose that for all $X \in N$, $\mathcal{L}_G(X) \neq \emptyset$. For each production $X \longrightarrow \omega_1 X_1 \omega_2 \ldots \omega_n X_n \omega_{n+1} \in P$, there are then $u_1, \ldots, u_n \in \mathcal{T}_\Gamma$ such that $\omega_1 u_1 \omega_2 \ldots \omega_n u_n \omega_{n+1} \in \mathcal{T}_\Gamma$. Using Lemma 3.1, $\rho(\omega_1 u_1 \omega_2 \ldots \omega_n u_n \omega_{n+1}) = \rho(\omega_1 \cdots \omega_{n+1})$ and then $\omega_1 \cdots \omega_{n+1} \in \mathcal{T}_\Gamma$.

For $(3 \Rightarrow 1)$, we consider a context-free grammar $G = (N, \widehat{\Gamma}, P, S)$ such that for all $X \longrightarrow \Omega \in P$: $\pi_{\widehat{\Gamma}}(\Omega) \in \mathcal{T}_\Gamma$. We transform G into an equivalent ε-CFG by applying the following process that transform iteratively each production $p : X \longrightarrow \Omega$, according to the Decomposition Property over $v_p = \pi_{\widehat{\Gamma}}(\Omega)$:

1. if $v_p = \varepsilon$ (that is, $\Omega \in N^*$), or $\Omega = \alpha Y \bar{\alpha}$, $Y \in N$, $\alpha \in \widehat{\Gamma}$, do not modify p;
2. else, if v_p is decomposable in $v_p = v_1 v_2$, with $v_1, v_2 \neq \varepsilon$ and $v_1, v_2 \in \mathcal{T}_\Gamma$, we choose a decomposition $\Omega = \Omega_1\Omega_2$ such that $\pi_{\widehat{\Gamma}}(\Omega_1) = v_1$ and $\pi_{\widehat{\Gamma}}(\Omega_2) = v_2$ and we replace p by the productions $X \longrightarrow Y_1 Y_2$, $Y_1 \longrightarrow \Omega_1$ and $Y_2 \longrightarrow \Omega_2$, where Y_1, Y_2 are fresh nonterminal symbols;
3. else, $v_p = \alpha v \bar{\alpha}$, with $v \in \mathcal{T}_\Gamma$ and $\alpha \in \widehat{\Gamma}$. In this case Ω can be decomposed as follows: $\Omega = \Omega_1 \alpha \Omega_2 \bar{\alpha} \Omega_3$, with $\Omega_1, \Omega_3 \in N^*$, and $\pi_{\widehat{\Gamma}}(\Omega_2) = v$. We replace then p by the productions $X \longrightarrow Y_1 Y_2 Y_3$, $Y_1 \longrightarrow \Omega_1$, $Y_2 \longrightarrow \alpha Z \bar{\alpha}$, $Y_3 \longrightarrow \Omega_3$ and $Z \longrightarrow \Omega_2$ where Y_1, Y_2, Y_3 and Z are fresh nonterminal symbols.

Remark that each step of the iteration preserves the property that for all $X \longrightarrow \Omega \in P$: $\pi_{\widehat{\Gamma}}(\Omega) \in \mathcal{T}_\Gamma$. Obviously, the process terminates and gives an ε-CFG equivalent to G. □

3.3. Closure properties

In this part we prove that ε-reducible context-free languages enjoy similar closures properties than context-free languages. The main difference being that, of course, ε-CFLs are not closed under homomorphism, nor inverse homomorphisms. We will however introduce a class of homomorphisms that preserves the family of ε-CFLs.

Proposition 3.6. *The class of ε-CFLs is closed under union, intersection with a regular language, concatenation and Kleene star.*

Proof. The class of ε-CFLs is obviously closed under union, concatenation and Kleene star. Only the closure under intersection with a regular language remains to prove.

Consider a language L generated by an ε-CFG $G = (N, \widehat{\Gamma}, P, S)$ and a regular language R. Since R is rational, there is a finite monoid M, a monoid morphism $\mu : \widehat{\Gamma}^* \to M$ 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 every $m \in H$.

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

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

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

An immediate consequence of Proposition 3.6 is that every regular language included in $\mathcal{T}_\Gamma$ is ε-reducible since it can be written as the intersection of $\mathcal{T}_\Gamma$ with itself.

Lemma 3.7. *Every regular language included in $\mathcal{T}_\Gamma$ is an ε-CFL.*

The family of ε-CFLs is obviously not closed under homomorphisms – at least, not without any restriction on the applied homomorphisms. In what follows, we introduce a class of homomorphisms that preserve the “ε-reducibility”.

Definition 3.8. A homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$ is said to be *ε-safe* if for all $u \in \widehat{\Sigma}^*$, $\rho(u) = \varepsilon$ implies $\rho(g(u)) = \varepsilon$.

In other words, g is ε-safe iff for all $\alpha \in \widehat{\Sigma}$, $\rho(g(\bar{\alpha})) = \rho(\overline{g(\alpha)})$.

Example 3.9. The homomorphism $g : \alpha \in \Gamma \mapsto \alpha\bar{\alpha};\ \bar{\alpha} \in \bar{\Gamma} \mapsto \varepsilon$ is ε-safe. □

Lemma 3.10. *For every ε-safe homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$: $g(\mathcal{T}_\Sigma) \subseteq \mathcal{T}_\Gamma$.*

Proof. By a trivial induction on the structure of words in $\mathcal{T}_\Sigma$. □

A homomorphism $h : A^* \to B^*$ is said to be *alphabetic* if for every $a \in A$, $h(a) \in B \cup \{\varepsilon\}$.

Obviously, if an ε-safe homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$ is alphabetic, then for all $\alpha \in \widehat{\Sigma}$: $g(\bar{\alpha}) = \overline{g(\alpha)}$. Therefore, when defining alphabetic ε-safe homomorphisms, we will generally just mention their mapping from the positive alphabet *(Σ in this case)*.

Proposition 3.11. *The class of ε-CFLs is closed under ε-safe homomorphism.*

Proof. Consider a language L generated by an ε-CFG $G = (N, \widehat{\Gamma}, P, S)$ and a ε-safe homomorphism $g : \widehat{\Gamma}^* \to \widehat{\Sigma}^*$. By replacing every production $X \longrightarrow u\Omega\bar{u}$ (with $\Omega \in N^*$) by $X \longrightarrow g(u)\Omega g(\bar{u})$, we obtain a context-free grammar generating $g(L)$ and such that every production $X \longrightarrow \Omega$ satisfies $\pi_{\widehat{\Sigma}}(\Omega) \in \mathcal{T}_\Sigma$ (from Lemma 3.10). According to Proposition 3.5, $g(L)$ is ε-reducible. □

Definition 3.12 *(Quotient).* Given a language $L \subseteq \Sigma^*$ and a word $w \in \Sigma^*$, we define:

- the right quotient of L by w: $L \cdot w^{-1} = \{u \mid uw \in L\}$;
- the left quotient of L by w: $w^{-1} \cdot L = \{u \mid wu \in L\}$.

Context-free languages are closed under quotient by a word w. It is also true for ε-CFLs as soon as $\rho(w) = \varepsilon$.

Proposition 3.13. *Let $L \in ECFL(\Gamma)$. For every $w \in \mathcal{T}_\Gamma$: $L \cdot w^{-1}$ and $w^{-1} \cdot L$ belong to $ECFL(\Gamma)$.*

Proof. We present the proof for the right product, the left product case is symmetrical. Suppose L to be generated by an ε-CFG $G = (N, \widehat{\Gamma}, P, S)$. We transform G in an equivalent grammar under Chomsky normal form $G = (N', \widehat{\Gamma}, P, S)$: for all $\alpha \in \widehat{\Gamma}$, we introduce a new nonterminal symbol X_α, we replace α by X_α in each production and we add productions $X_\alpha \longrightarrow \alpha$.

We define now R_w to be the set of all $\Omega \in N^*$ such that $S \xrightarrow{*}_{G'} \Omega' \longrightarrow_{G'} \Omega w$ by a **rightmost** derivation such that w is not a suffix of Ω' (this means that the last step of the derivation has derived the first letter of w). It is folklore that R_w is regular. In addition, from definition of R_w, $\{u \in \Sigma^* \mid \Omega \xrightarrow{*}_{G'} u, \Omega \in R_w\} = L \cdot w^{-1}$.

We consider now the morphism $g : N' \to N \cup \widehat{\Gamma}$ that replace all nonterminals X_α by α, and define $Q = g(R_w)$. Obviously we get:

Claim 1. *Q is regular and $L \cdot w^{-1} = \{u \in \widehat{\Gamma}^* \mid \Omega \xrightarrow{*}_G u, \Omega \in Q\}$.*

Then, for every $\Omega \in Q$, and every $u \in \widehat{\Gamma}^*$ such that $\Omega \xrightarrow{*}_G u$, $uw \in \mathcal{T}_\Gamma$. It follows that $u \in \mathcal{T}_\Gamma$, and since G is an ε-CFG, $\pi_{\widehat{\Gamma}}(\Omega) \in \mathcal{T}_\Gamma$. In other words:

Claim 2. $\pi_{\widehat{\Gamma}}(Q) \subseteq \mathcal{T}_\Gamma$.

Consider now the morphism $\mu : (N \cup \widehat{\Gamma})^* \to (\widehat{N} \cup \widehat{\Gamma})^*$ defined by $X \in N \mapsto X\bar{X}$ and $\alpha \in \widehat{\Gamma} \mapsto \alpha$. From Claims 1 and 2, the language $\mu(Q)$ is regular and included in $\mathcal{T}_{N \cup \Gamma}$. It is then also an ε-CFL from Lemma 3.7.

Let $G_w = (N_w, \widehat{N} \cup \widehat{\Gamma}, P_w, S_w)$ be an ε-CFG generating $\mu(Q)$, we define the context free grammar $G' = (N', \widehat{\Gamma}, P_w \cup P \cup P', S_w)$, where $N' = N \cup N_w \cup \bar{N}$ and $P' = \{\bar{X} \longrightarrow \varepsilon, X \in N\}$.

From construction, $S_w \xrightarrow{*}_{G'} u$ iff there is $\Omega \in Q$ such that $\Omega \xrightarrow{*}_G u$ iff (from Claim 1) $u \in L \cdot w^{-1}$. Then:

Claim 3. $\mathcal{L}_{G'} = L \cdot w^{-1}$.

We conclude by emphasizing that for any production $X \longrightarrow \Omega \in P'$, $\pi_{\widehat{\Gamma}}(\Omega) \in \mathcal{T}_\Gamma$ and hence from Proposition 3.5, $\mathcal{L}_{G'}$ is an ε-CFL. □

Another well known property of context-free language is the closure under reversal. Given a word $w = \alpha_1 \cdots \alpha_n$, the reversal of w is $w^R = \alpha_n \cdots \alpha_1$. The reversal of a language L is $L^R = \{w^R \mid w \in L\}$.

Proposition 3.14. *The class of ε-CFLs is closed under reversal.*

Proof. The proof is similar to that of closure under reversal of context-free languages. Given an ε-CFG $G = (N, \widehat{\Gamma}, P, S)$, we define $G' = (N, \widehat{\Gamma}, P^R, S)$ where P^R is the set of productions $X \longrightarrow \Omega^R$, for $X \longrightarrow \Omega \in P$. Clearly, G' is an ε-CFG and we can check by induction on length of derivations that for all $X \in N$, $X \xrightarrow{*}_G w \in \widehat{\Gamma}^*$ iff $X \xrightarrow{*}_{G'} w^R$. It follows that $\mathcal{L}_{G'} = {\mathcal{L}_G}^R$. □

We show now that every context-free language L can be mapped to an ε-CFL by a homomorphism. This property will be used in several proofs.

Lemma 3.15. *Let $L \subseteq \Sigma^*$ be a context-free language and $\mu : \Sigma^* \to \widehat{\Gamma}^*$ be a homomorphism such that: for all $\alpha \in \Sigma$, $\mu(\alpha) \in \mathcal{T}_\Gamma$. The language $\mu(L)$ is an ε-CFL.*

Proof. Suppose L to be generated by a context-free grammar $G = (N, \Sigma, P, S)$ in Chomsky normal form. Replacing every production $X \longrightarrow u$, $u \in \Sigma \cup \{\varepsilon\}$ by $X \longrightarrow \mu(u)$, we get a context-free grammar generating $\mu(L)$ and in which every nonterminal X generates only words in $\mathcal{T}_\Gamma$, that is, from Proposition 3.5, $\mu(L)$ is an ε-CFL. □

Proposition 3.16. *For all Γ, the class $ECFL(\Gamma)$ is not closed under intersection nor complement in $\mathcal{T}_\Gamma$.*

Proof. Suppose that $E = L_1 \cap L_2$ where $L_1, L_2 \in \Gamma^*$ are context-free languages, and E is not context-free (such languages exist since context-free languages are not closed under intersection). Let $\mu : \Gamma^* \to \widehat{\Gamma}^*$ be the morphism defined by $\alpha \mapsto \alpha\bar{\alpha}$, and π_Γ be the projection of $\widehat{\Gamma}^*$ into Γ^*; E can be written:

$$E = \pi_\Gamma(\mu(L_1) \cap \mu(L_2)),$$

since μ is injective and $\pi_\Gamma \circ \mu$ is the identity map.

The sets $\mu(L_1)$ and $\mu(L_2)$ are ε-CFLs (see Lemma 3.15) but $\mu(L_1) \cap \mu(L_2)$ is not context-free since context-free languages are closed under homomorphism. It follows that $\mu(L_1) \cap \mu(L_2)$ is not an ε-CFL and then $ECFL(\Gamma)$ is not closed under intersection.

Now, let L_1, L_2 be two ε-CFLs. Since $L_1, L_2 \subseteq \mathcal{T}_\Gamma$, we have:

$$L_1 \cap L_2 = \mathcal{T}_\Gamma - ((\mathcal{T}_\Gamma - L_1) \cup (\mathcal{T}_\Gamma - L_2)).$$

Since $ECFL(\Gamma)$ is closed under union, if it is closed under complement in $\mathcal{T}_\Gamma$, it is also closed under intersection. Then $ECFL(\Gamma)$ is not closed under complement in $\mathcal{T}_\Gamma$. □

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

3.4. 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 $R \cap \mathcal{D}_\Gamma$ is an ε-CFL. However, if we restrict h to be ε-safe, we generate exactly the family of ε-CFLs.

Proposition 3.17. *For all $L \in ECFL(\Gamma)$, there is an alphabet Σ, an alphabetic ε-safe homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$, and a regular language $R \subseteq \widehat{\Sigma}^*$ such that $L = g(R \cap \mathcal{D}_\Sigma)$.*

Proof. We use a slight adaptation of the proof of the non-erasing variant of the Chomsky–Schützenberger theorem given in [20]. Suppose L to be generated by an ε-CFG $G = (N, \widehat{\Gamma}, S, P)$ such that all productions have the form $X \longrightarrow \alpha_\varepsilon \bar{\alpha_\varepsilon}$, or $X \longrightarrow YZ$ where symbols Y, Z are distinct and $\alpha_\varepsilon \in \widehat{\Gamma} \cup \{\varepsilon\}$. This condition is not restrictive since such a grammar can easily be obtained.

We build from G the context-free grammar $G' = (N', \widehat{\Sigma}, P', S')$ where

- $N' = \{X_p \mid X \in N, p \in P \cup \{\#\}\}$;
- Σ is the set of pairs of productions (p_0, p_1) such that p_1 *is compatible with* p_0, that is
$$\Sigma = \{(p_0, p_1) \mid p_0 = \# \text{ or } p_0 = X \longrightarrow \Omega_1 Y \Omega_2 \text{ and } p_1 = Y \longrightarrow \Omega_3\};$$
- the set of productions P' consists of all $X_{p_0} \longrightarrow (p_0, p_1) X_{1,p_1} \cdots X_{n,p_1} \overline{(p_0, p_1)}$ such that $(p_0, p_1) \in \Sigma$ and $p_1 = X \longrightarrow \alpha_\varepsilon X_1 \cdots X_n \bar{\alpha_\varepsilon} \in P$.

Consider now the alphabetic ε-safe homomorphism $g : \widehat{\Sigma}^* \to \Gamma^*$ defined as follows:

$$\text{for all } (p_0, p_1) \in \Sigma, \text{ with } p_1 = X \longrightarrow \alpha_\varepsilon X_1 \cdots X_n \bar{\alpha_\varepsilon},\ g(p_0, p_1) = \alpha_\varepsilon.$$

We have clearly:

Claim 1. *$\mathcal{L}_{G'} \subseteq \mathcal{D}_\Sigma$ and $\mathcal{L}_G = g(\mathcal{L}_{G'})$.*

Now, we construct a regular language $R \subseteq \Sigma^*$ such that $\mathcal{L}_{G'} = R \cap \mathcal{D}_\Sigma$. First, we define R_0 as the set of all words in $\widehat{\Sigma}^*$ such that each factor of length 2 is in one of the following forms:

- $(p_0, p_1)(p_1, p_2)$ for $(p_1 = X \longrightarrow \alpha_\varepsilon Y \bar{\alpha_\varepsilon}$ or $p_1 = X \longrightarrow YZ)$ and $p_2 = Y \longrightarrow \Omega$;
- $(p_0, p_1)\overline{(p_0, p_1)}$ for $p_1 = X \longrightarrow \alpha_\varepsilon \bar{\alpha_\varepsilon}$;
- $\overline{(p_1, p_2)}(p_1, p_3)$ for $p_1 = X \longrightarrow YZ$ and $p_2 = Y \longrightarrow \Omega$ and $p_3 = Z \longrightarrow \Omega'$;
- $\overline{(p_1, p_3)}\,\overline{(p_0, p_1)}$ for $(p_1 = X \longrightarrow YZ$ or $p_1 = X \longrightarrow \alpha_\varepsilon Z \bar{\alpha_\varepsilon})$ and $p_3 = Z \longrightarrow \Omega$.

The set R_0 is a regular set expressing a necessary condition on derivations, then $\mathcal{L}_{G'} \subseteq R_0$, and so, $\mathcal{L}_{G'} \subseteq R_0 \cap \mathcal{D}_\Sigma$.

Claim 2. *For all words $u = (p_0, p_1) v \overline{(p_0, p_1)} \in R_0 \cap \mathcal{D}_\Sigma$, there is a derivation of u in G' starting with the production p_1.*

This can be checked by induction on the length of u.

Hence, let

$$R = R_0 \cap \bigcup_{p = S \longrightarrow \Omega \in P} (\#, p) \widehat{\Sigma}^* \overline{(\#, p)},$$

we have $R \cap \mathcal{D}_\Sigma = \mathcal{L}_{G'}$, and then $\mathcal{L}_G = g(R \cap \mathcal{D}_\Sigma)$. □

Theorem 3.18. *Let $L \subseteq \widehat{\Gamma}^*$, the following assertions are equivalent:*

1. *L is an ε-CFL;*
2. *there is an alphabet Σ, an ε-safe homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$, and a regular set $R \subseteq \widehat{\Sigma}^*$ such that $L = g(R \cap \mathcal{D}_\Sigma)$;*
3. *there is an alphabet Σ, an alphabetic ε-safe homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$, and a regular set $R \subseteq \widehat{\Sigma}^*$ such that $L = g(R \cap \mathcal{D}_\Sigma)$;*

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

Proof. ($3 \Rightarrow 2$) is obvious; ($2 \Rightarrow 1$) is a straightforward consequence of Propositions 3.6, 3.11 and the fact that $\mathcal{D}_\Sigma$ is an ε-CFL; ($1 \Rightarrow 3$) is stated Proposition 3.17. □

We conclude by emphasizing that Theorem 3.18, Proposition 3.6 and Proposition 3.11 provide another characterization of the class of ε-CFLs:

Corollary 3.19. *The family of ε-CFLs is the least family of languages that contains the Dyck languages and is closed under union, intersection with a regular language, ε-safe homomorphisms, concatenation and Kleene star.*

3.5. Comparison with two-sided Dyck context-free languages

Let us now compare $ECFL(\Gamma)$ with the class of context-free languages included in $\mathcal{T}_\Gamma$. We show these two classes to be distinct using a pumping lemma.

Lemma 3.20. *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 generating 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 $\rho(w) = \varepsilon$. □

Definition 3.21. A language $L \subseteq \widehat{\Gamma}^*$ is a Two-sided Context-Free Language (TCFL) if it is context-free and included in $\mathcal{T}_\Gamma$.

Proposition 3.22. *There is a TCFL which is not an ε-CFL.*

Proof. Let us consider the TCFL $L = \{(a\bar{a})^n b(a\bar{a})^n \bar{b} \mid n \geq 0\}$. Applying Lemma 3.20 to L, we consider a word $s = (a\bar{a})^m b(a\bar{a})^m \bar{b}$ such that $|s| \geq p$. There is then a decomposition $s = uvwxy$ such that

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 $n \geq 0$.

Clearly, to satisfy conditions 3 and 4, it is necessary that $v = (\alpha\bar{\alpha})^i$ and $x = (\beta\bar{\beta})^i$ for some $1 \leq i \leq m$ and some $\alpha, \beta \in \{a, \bar{a}\}$, and there are $w_1, w_2 \in \{a, \bar{a}\}^*$ such that $w = w_1 b w_2$. It follows that $\rho(w) \neq \varepsilon$, contradicting the condition 1. □

3.6. Decision problems

Our main decidability result is that it is undecidable to know whether a context-free language L is an ε-CFL. In particular, this problem is undecidable even if L is assumed to be a TCFL, since one can decide if a context-free language is a TCFL [21], and a context-free language is an ε-CFL only if it is a TCFL.

The proof is inspired by the Greibach Theorem that gives a general method to deal with this kind of problems, and allows for example to prove that it is undecidable to know if a context-free language is rational (see for example [22]).

We start by showing that is undecidable to determine if a language $L \in ECFL(\Gamma)$ is equal to $\mathcal{T}_\Gamma$.

Proposition 3.23. *The problem "$L = \mathcal{T}_\Gamma$?" is undecidable for $L \in ECFL(\Gamma)$.*

Proof. Let us consider the morphism $\mu : \Gamma^* \to \widehat{\Gamma}^*$ defined by $a \mapsto a\bar{a}$. The first step of the proof is to note that every word $w \in \mathcal{T}_\Gamma$ admits a (unique) decomposition $w = w_0 \mu(v_1) w_1 \cdots \mu(v_n) w_n$, $n \geq 0$ where:

- for all $0 \in [0, n]$, $w_i \in \widehat{\Gamma}^*$ and for all $a \in \Gamma$, w_i does not contain occurrences of $a\bar{a}$;
- for all $i \in [1, n]$, $v_i \in \Gamma^+$;
- for all $i \in [1, n-1]$ $w_i \neq \varepsilon$.

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

To get this decomposition, it suffices to choose $v_1, \dots, v_n$ so that $|v_1| + \cdots + |v_n|$ is maximal. Let $K \subseteq \Gamma^+$. If for all $i \in [1, n]$, $v_i \in K$, w is said to be K-decomposable.

Claim 1. *If $K \subseteq \Gamma^+$ is context-free, the set $L_K \subseteq \mathfrak{T}_\Gamma$ of K-decomposable words is an ε-CFL.*

Proof of Claim 1. Let # be a symbol that does not belong to Γ, and

- R be the set of words over $\widehat{\Gamma} \cup \widehat{\{\#\}}$ that contain no factor $a\bar{a}$, $a \in \Gamma$,
- $R' = \widehat{A}^*(\#\bar{\#}\widehat{A}^+)^*\#\bar{\#}\widehat{A}^*$.

Languages R and R' and regular, then the language $M = \mathfrak{T}_{\Gamma \cup \{\#\}} \cap R \cap R'$ is an ε-CFL by Proposition 3.6 and is equal to $L_{\{\#\}}$.

According to Lemma 3.15, $\mu(K)$ is an ε-CFL that we assume to be generated by the ε-CFG $G_\# = (N, \widehat{\Gamma}, P, S_\#)$. We transform the ε-CFG generating M as follows: in every production, we replace $\bar{\#}$ by ε, and # by $S_\#$, then we add the set of productions P. The grammar thus constructed is an ε-CFG that generate L_K. □

Note that for all language $K \subseteq \Gamma^*$, L_K contains at least the set of words over $\widehat{\Gamma}$ that contain no factor $a\bar{a}$, $a \in \Gamma$.

Claim 2. *$L_K = \mathfrak{T}_\Gamma$ iff $K = \Gamma^+$.*

Proof of Claim 2. Suppose that $K \neq \Gamma^+$: there is $u \in \Gamma^+$ such that $u \notin K$, then $\mu(u) \notin L_K$ and hence $L_K \neq \mathfrak{T}_\Gamma$. Conversely, given $w \in \mathfrak{T}_\Gamma$, if $w \notin L_K$, then w is not K-decomposable and contains at least a factor of the form $a\bar{a}$, for some $a \in \Gamma$. This means that the decomposition of w is $w = w_0\mu(v_1)w_1 \cdots \mu(v_n)w_n$, with $n \geq 1$ and there is $i \in [1, n]$ such that $v_i \notin K$. In particular, $K \neq \Gamma^+$. □

The problem to know whether a context-free language $K \subseteq \Gamma^*$ satisfies $K = \Gamma^*$ is undecidable [22, Thm 8.10], then it is easy to see that the problem to know whether a context-free language $K \subseteq \Gamma^+$ satisfies $K = \Sigma^+$ is too. We conclude the problem $L = \mathfrak{T}_\Gamma$? for $L \in ECFL(\Gamma)$ to be undecidable. □

Theorem 3.24 *([21]). Given a context-free language L, one can decide in polynomial time whether L is a TCFL.*

Theorem 3.25. *The problem to know whether a context-free language L is an ε-CFL is undecidable. The problem is also undecidable if L is a TCFL.*

Proof. Let $L_1 \subseteq \widehat{\Sigma}^*$ be an ε-CFL, and $L_0 \subseteq \widehat{\Sigma}^*$ be a TCFL which is not an ε-CFL (such a language exists from Proposition 3.22). Let $L = L_0\#\bar{\#}\mathfrak{T}_\Sigma \cup \mathfrak{T}_\Sigma\#\bar{\#}L_1$ where # is a new symbol that does not belong to Σ. From closure properties of context-free languages, L is a context-free language, and since $L \in \mathfrak{T}_{\Sigma \cup \{\#\}}$, it follows that L is a TCFL.

Claim. *L is an ε-CFL iff $L_1 = \mathfrak{T}_\Sigma$.*

Proof of Claim. Clearly, if $L_1 = \mathfrak{T}_\Sigma$, then $L = \mathfrak{T}_\Sigma\#\bar{\#}\mathfrak{T}_\Sigma$ which is an ε-CFL. Let us suppose that $L_1 \neq \mathfrak{T}_\Sigma$. Then, there is $v \in \mathfrak{T}_\Sigma$ such that $v \notin L_1$, and hence $L \cap \mathfrak{T}_\Sigma\#\bar{\#}v = L_0\#\bar{\#}v$. The set $L_0\#\bar{\#}v$ is not an ε-CFL by Proposition 3.13, then $L \cap (\mathfrak{T}_\Sigma\#\bar{\#}v)$ is not an ε-CFL. Hence L is not an ε-CFL since $L \cap (\mathfrak{T}_\Sigma\#\bar{\#}v) = L \cap (\widehat{\Sigma}^*\#\bar{\#}v)$ and from Proposition 3.6, ε-CFLs are closed under intersection with regular languages. □

We have then reduced the problem to know whether L is an ε-CFL, to the problem to know if $L_1 = \mathfrak{T}_\Sigma$ which is undecidable by Proposition 3.23. □

4. 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 whose domains are ε-CFLs. We give a Nivat-like presentation of those transductions.

4.1. Transductions

We briefly introduce rational and context-free transductions. The reader can refer to [23] 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 [24] 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 establishes 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 [23, page 62], a (Γ, Σ)-transduction τ 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: 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, [23, page 62] they are not closed under composition and classes of languages are usually not closed under them. For example, every context-free language is the image of a regular language, and every recursively enumerable language is the image of a context-free language.

4.2. ε-Reducible context-free transductions and transducers

Definition 4.1. An ε**-Reducible Context-Free Transduction Grammar** (ε-CFTG) is a context-free transduction grammar $G = (N, \widehat{\Gamma}, \Sigma, S, P)$ in which every production is of 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 4.2. Let $G = (N, \{\alpha, \beta, \bar{\alpha}, \bar{\beta}\}, \{a\}, S, P)$ be the ε-CFTG whose productions are:

$$\begin{array}{lll} S \longrightarrow (\beta, \varepsilon)X(\bar{\beta}, \varepsilon), & X \longrightarrow (\alpha, \varepsilon)X(\bar{\alpha}, \varepsilon), & X \longrightarrow (\varepsilon, \varepsilon)Y(\varepsilon, \varepsilon), \\ Y \longrightarrow (\bar{\alpha}, a)YZ(\alpha, \varepsilon), & Z \longrightarrow (\bar{\alpha}, a)Z(\alpha, a), & Y \longrightarrow (\bar{\beta}, \varepsilon)(\beta, \varepsilon),\ Z \longrightarrow (\bar{\beta}, \varepsilon)(\bar{\beta}, \varepsilon). \end{array}$$

Let τ be the transduction generated by G. The domain of τ is the ε-CFL given in Example 3.4 and one can easily check that

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

By a proof similar to that of Proposition 3.5, we get:

Proposition 4.3. *Given $\tau \subseteq \widehat{\Gamma}^* \times \Sigma^*$, the following properties are equivalent:*

1. *L is an ε-CFT;*
2. *there is a context-free transduction grammar $G = (N, \widehat{\Gamma}, \Sigma, P, S)$ such that $\mathcal{T}_G = \tau$ and for all $X \in N$: if $X \xrightarrow{*}_G (\omega, u)$, then $\omega \in \mathfrak{T}_\Gamma$.*

To ease the proofs, we will often use grammars under a simpler form, thanks to the following property.

Lemma 4.4. *Every ε-CFT can be generated by an ε-CFTG in which every production is in the form*

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

An ε-CFTG under this form is said to be **standard**. □

Proof. Let $G = (N, \widehat{\Gamma}, \Sigma, S, P)$ be an ε-CFTG. One simply has to replace every production $X \longrightarrow (\omega, u)\Omega(\overline{\omega}, v)$, with $\omega = \alpha_1 \cdots \alpha_n$ and $\Omega \in N^*$, by $X \longrightarrow (\varepsilon, u)Y_1(\varepsilon, v)$ and introduce the new productions $Y_i \longrightarrow (\alpha_i, \varepsilon)Y_{i+1}(\bar{\alpha}_i, \varepsilon)$ for $i \in \{1, \ldots, n-1\}$ and $Y_n \longrightarrow (\alpha_n, \varepsilon)\Omega(\bar{\alpha}_n, \varepsilon)$. □

Theorem 4.5. *Given a $(\widehat{\Gamma}, \Sigma)$-transduction τ, the following assertions are equivalent:*

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

1. *τ is an ε-CFT.*
2. *There is an alphabet Δ, an ε-CFL $X \subseteq \widehat{\Delta}^*$, an ε-safe (alphabetic) 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, an ε-safe (alphabetic) 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) According to Lemma 4.4), we suppose τ to be generated by a standard ε-CFTG $G = (N, \widehat{\Gamma}, \Sigma, S, P)$. We define the grammar $G' = (N, \widehat{\Delta}, S, P')$ where $\Delta = P$, and

$$P' = \{X \longrightarrow p\Omega\bar{p} \mid p = X \longrightarrow (\alpha_\varepsilon, v)\Omega(\bar{\alpha}_\varepsilon, w) \in P\}.$$

Now, let $h : \widehat{\Delta}^* \to A^*$ be the homomorphism defined by

$$\text{for all } p = X \longrightarrow (\alpha_\varepsilon, v)\Omega(\bar{\alpha}_\varepsilon, w)\text{:}\;\; p \mapsto v,\; \bar{p} \mapsto w,$$

and $g : \widehat{\Delta}^* \to \widehat{\Gamma}^*$ be the ε-safe alphabetic homomorphism defined by

$$\text{for all } p = X \longrightarrow (\alpha_\varepsilon, v)\Omega(\bar{\alpha}_\varepsilon, w)\text{:}\;\; p \mapsto \alpha_\varepsilon,\; \bar{p} \mapsto \bar{\alpha_\varepsilon}.$$

Clearly we G' is an ε-CFG and $\mathcal{T}_G = \{(g(u), h(u)) \mid u \in \mathcal{L}(G')\}$.

(2 $\Rightarrow$ 3) Suppose that $\tau = \{(g(u), h(u)) \mid u \in X\}$ where X is an ε-CFL and g is (alphabetic) ε-safe. From Theorem 3.18, there is an alphabet Σ, a regular set $R \subseteq \widehat{\Sigma}^*$, and a ε-safe alphabetic homomorphism $g' : \widehat{\Sigma}^* \to \widehat{\Delta}^*$ such that $X = g'(R \cap \mathcal{D}_\Delta)$. The homomorphism $g \circ g'$ is (alphabetic) ε-safe 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 ε-safe. From Proposition 3.6, $R \cap \mathcal{D}_\Delta$ is an ε-CFL that we assume to be generated by the ε-CFG $G = (N, \widehat{\Delta}, P, S)$. Then τ is generated by the context-free grammar $G' = (N, \widehat{\Gamma}, \widehat{\Sigma}, P', S)$ where

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

From Proposition 3.5, the domain of $\mathcal{T}_{G'}$ is an ε-CFL, and so τ is an ε-CFT. □

It is clear that every context-free language L can be obtained by applying an ε-CFT to a regular language (for instance using the transduction $\{\varepsilon\} \times L$). We will see (Theorem 5.6) 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 4.6. *For every recursively enumerable language E, there is an ε-CFT τ, and an ε-CFL Z such that $E = \tau(Z)$.*

Proof. Suppose that $E \subseteq \Sigma^*$. 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 ε-safe 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 4.5, $E = \tau(Z)$, where τ is an ε-CFT. We conclude by noting that from Lemma 3.15, Z is an ε-CFL. □

5. Characterizations of indexed languages

We relate now 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.

5.1. Indexed grammars and languages

Introduced by Aho [3], 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 LIFO process: during a step of derivation of X^ω, it is possible to add a symbol in head of 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

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

Indexes are denoted as *superscript*, and we will often 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

$$\Omega_1 A^{\eta\omega} \Omega_2 \longrightarrow_{\mathcal{I}} \Omega_1 u_0 B_1{}^{\eta_1 \omega} \cdots B_n{}^{\eta_1 \omega} u_n \Omega_2$$
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 5.1. Let us consider the following indexed grammar $\mathcal{I} = (N, I, \Sigma, S, P)$ with $N = \{S, X, A, B, C\}$, $I = \{\beta, \alpha\}$, $\Sigma = \{a, b, c\}$ 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 $\{a^n b^n c^n, n \geq 0\}$. □

5.2. Characterizations of indexed languages by context-free transductions

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

Theorem 5.2 *([22]). For every indexed language L, there is an indexed grammar (N, I, T, S, P) generating L in which every production is in either one of the following forms.*

$$(1)\ X \longrightarrow YZ, \quad (2)\ X \longrightarrow Y^{\alpha}, \quad (3)\ X^{\alpha} \longrightarrow Y,$$
$$\text{or } (4)\ X \longrightarrow u$$
with $X, Y, Z \in N$, $a \in T$, $u \in T^*$ and $\alpha \in I$.

Corollary 5.3. *Every indexed language can be generated by an indexed grammar in which every production is in one of these forms*

$$(1)\ X_0 \longrightarrow uX_1{}^{\alpha} \cdots X_n{}^{\alpha} v, \quad \text{or } (2)\ X_0{}^{\alpha} \longrightarrow uX_1 \cdots X_n v$$
with $n \geq 0$, $X_i \in N$, $\alpha \in I \cup \{\varepsilon\}$ and $u, v \in T^*$.

An indexed grammar under this form is said to be **standard***.*

Definition 5.4. Let us consider the mapping φ that maps a standard indexed grammar $\mathcal{I} = (N, I, \Sigma, P, S)$ into a standard ε-CFTG $\varphi(\mathcal{I}) = (N, \widehat{I}, \Sigma, \varphi(P), S)$ by transforming every production

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

Fact. The transformation φ is bijective.

Lemma 5.5. *Let $\mathcal{I} = (N, I, \Sigma, P, S)$ be a standard indexed grammar.*

There is a derivation

$$S \xrightarrow{*}_{\mathcal{I}} v_1 Y_1{}^{w_1} v_2 Y_2{}^{w_2} \cdots Y_n{}^{w_n} v_{n+1}$$

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

iff there is a derivation

$$S \xrightarrow{*}_{\varphi(\mathcal{I})} (u_1, v_1)Y_1(u_2, v_2)Y_2 \cdots Y_n(u_{n+1}, v_{n+1}) \textit{ such that}$$
$$u_1 \cdots u_{n+1} \in \mathcal{D}_I, \textit{ and } \rho(u_1 \cdots u_i) = w_i^R, \textit{for all } i \in [1, n].$$

Here w_i^R denotes the reverse of w_i.

Proof. This is proved by induction over the steps of derivations and is trivially true derivations of length 0. Furthermore, since the derivation rules of indexed grammars, and of context-free grammar are confluent, we need only to consider leftmost derivations for our induction. Let us prove the induction step. Suppose a derivation

$$S \xrightarrow{*}_{\mathcal{I}} v_1 {Y_1}^{w_1} v_2 {Y_2}^{w_2} \cdots {Y_n}^{w_n} v_{n+1}$$

in $\mathcal{I}$ and a derivation

$$S \xrightarrow{*}_{\varphi(\mathcal{I})} (u_1, v_1)Y_1(u_2, v_2)Y_2 \cdots Y_n(u_{n+1}, v_{n+1})$$

in $\varphi(\mathcal{I})$ with $u_1 \cdots u_{n+1} \in \mathcal{D}_I$ and $\rho(u_1 \cdots u_i) = w_i^R$ for $i \in [1, n]$.

Because φ is bijective, there is a production $p = Y_1 \longrightarrow x{Z_1}^{\alpha} \cdots {Z_m}^{\alpha} y \in P$ iff there is a production $\varphi(p) = Y_1 \longrightarrow (\alpha, x)Z_1 \cdots Z_m(\bar{\alpha}, y) \in \varphi(P)$. By applying p and $\varphi(p)$, we obtain the derivations

$$S \xrightarrow{*}_{\mathcal{I}} v_1 x {Z_1}^{\alpha w_1} \cdots {Z_m}^{\alpha w_1} y v_2 {Y_2}^{w_2} \cdots {Y_n}^{w_n} v_{n+1}$$

in $\mathcal{I}$ and

$$S \xrightarrow{*}_{\varphi(\mathcal{I})} (u_1\alpha, v_1 x)Z_1 \cdots Z_m(\bar{\alpha}u_2, y v_2)Y_2 \cdots Y_n(u_{n+1}, v_{n+1})$$

in $\varphi(\mathcal{I})$. Furthermore, since $\alpha \in I \cup \{\varepsilon\}$, the word $u_1\alpha\bar{\alpha}u_2 \cdots u_{n+1}$ is in $\mathcal{D}_I$. Also, $\rho(u_1\alpha) = \rho(u_1)\alpha = (\alpha w_1)^R$ and for every $i \in \{2, \ldots, n\} : \rho(u_1\alpha\bar{\alpha} \cdots u_i) = \rho(u_1 \cdots u_i) = {w_i}^R$.

Likewise, there is a production $p = Y_1^{\alpha} \longrightarrow xZ_1 \cdots Z_m y \in P$ iff there is a production $\varphi(p) = Y_1 \longrightarrow (\bar{\alpha}, x)Z_1 \cdots Z_m(\alpha, y) \in \varphi(P)$. This time around, we need to observe two cases.

- If the production p can be applied to derivation in $\mathcal{I}$, it means that $w_1 = \alpha w_0$ and $\rho(u_1\bar{\alpha}) = w_0$ for some $w_0 \in I^*$. Also, the word $u_1\bar{\alpha}\alpha \cdots u_n$ still belongs to $\mathcal{D}_I$. By applying p and $\varphi(p)$ we obtain the derivations

  $$S \xrightarrow{*}_{\mathcal{I}} v_1 x {Z_1}^{\alpha w_1} \cdots {Z_m}^{\alpha w_1} y v_2 {Y_2}^{w_2} \cdots {Y_n}^{w_n} v_{n+1}$$

  in $\mathcal{I}$ and

  $$S \xrightarrow{*}_{\varphi(\mathcal{I})} (u_1\bar{\alpha}, v_1 x)Z_1 \cdots Z_m(\alpha u_2, y v_2)Y_2 \cdots Y_n(u_{n+1}, v_{n+1})$$

  in $\varphi(\mathcal{I})$ satisfying the induction hypothesis.
- If the production p cannot be applied to derivation in $\mathcal{I}$, it means that w_1 does not start with α and $\rho(u_1)$ does not end with α. Therefore, the word $u_1\bar{\alpha}\alpha u_2 \cdots u_{n+1}$ does not belong to $\mathcal{D}_I$. Applying $\varphi(p)$ to the derivation in $\varphi(\mathcal{I})$ would then result in a derivation

  $$S \xrightarrow{*}_{\varphi(\mathcal{I})} (u_1\bar{\alpha}, v_1 x)Z_1 \cdots Z_m(\alpha u_2, x v_2)Y_2 \cdots Y_n(u_{n+1}, v_{n+1})$$

  that does not satisfy the induction hypothesis. □

Theorem 5.6. *A language is indexed iff there is an ε-CFT τ and $k \geq 0$ such that $L = \tau(\mathcal{D}_k)$.*

Proof. ($\Rightarrow$) Let $\mathcal{I} = (N, I, \Sigma, P, S)$ be a standard indexed grammar. Lemma 5.5 implies that there is a terminal derivation $S \xrightarrow{*}_{\mathcal{I}} w$ in $\mathcal{I}$ iff there is a terminal derivation $S \xrightarrow{*}_{\varphi(\mathcal{I})} (u, w)$ in $\varphi(\mathcal{I})$ with $u \in \mathcal{D}_I$. Therefore, for every standard indexed grammar $\mathcal{I} = (N, I, \Sigma, P, S)$:

$$\mathcal{L}_{\mathcal{I}} = \tau_{\varphi(\mathcal{I})}(\mathcal{D}_I).$$

($\Leftarrow$) Conversely, since φ is bijective, for every reduced ε-CFTG $G = (N, \widehat{\Gamma}, \Sigma, P, S)$:

$$\tau_G(\mathcal{D}_\Gamma) = \mathcal{L}_{\varphi^{-1}(G)}.$$

Together with Lemma 4.4 and Corollary 5.3, this proves the theorem. □

Example 5.7. Let $\mathcal{I} = (N, I, \Sigma, S, P)$ be an indexed grammar with $N = \{S, X, Y, W, Z\}$, $I = \{\beta, \alpha\}$, $A = \{a\}$ and P consists of the rules:

$$\begin{array}{llll} S \longrightarrow X^{\beta}, & X \longrightarrow X^{\alpha}, & X \longrightarrow Y, & Y^{\alpha} \longrightarrow aYZ \\ Y^{\beta} \longrightarrow \varepsilon, & Z^{\alpha} \longrightarrow aZa, & Z^{\beta} \longrightarrow \varepsilon. & \end{array}$$

Initially defined in [25], the grammar $\mathcal{I}$ generates the language $L = \{a^{n^2} \mid n \geq 0\}$.

Applying the bijection φ defined above to $\mathcal{I}$, we get the ε-CFTG G given in Example 4.2 and generating the transduction

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

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

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

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

Corollary 5.8. *A language L is indexed if there is a homomorphism h, an ε-safe homomorphism g, a regular set R and $k, q \in \mathbb{N}$ such that*

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

This characterization is still true if g is alphabetic.

Proof. Direct from Theorem 5.6 and Theorem 4.5. □

For every positive integer k, the set $\mathcal{D}_k$ can be written as $\mu^{-1}(\mathcal{D}_2)$ where μ is the ε-safe and injective homomorphism that encodes a positive symbol a_i as 01^i0 and $\bar{a}_i$ as $\bar{0}\bar{1}^i\bar{0}$. Combined with the closure under composition of ε-safe homomorphisms, we obtain the following.

Corollary 5.9. *Let L be language. The following assertions are equivalent.*

1. *L is indexed;*
2. *there is a homomorphism h, an ε-safe homomorphism g, a regular set R and $k \in \mathbb{N}$ such that*

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

3. *there is an ε-CFT τ such that $L = \tau(\mathcal{D}_2)$.*

Another proof of the following theorem can be found in [15].

Theorem 5.10. *A language L is indexed iff there is an ε-CFL K, a homomorphism h, and an alphabet Δ such that*

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

Proof. ($\Rightarrow$) Let $L \subseteq A^*$ be an indexed language. From Theorem 5.6 and Theorem 4.5, there are alphabets Σ, Γ, an ε-CFL $K \subseteq \widehat{\Sigma}^*$, a homomorphism $h : \widehat{\Sigma}^* \to A^*$ and an ε-safe homomorphism $g : \widehat{\Sigma}^* \to \widehat{\Gamma}^*$ such that $L = h(K \cap g^{-1}(\mathcal{D}_\Gamma))$. We suppose that $\Gamma \cap A = \emptyset$ (otherwise, it suffices to work with a copy of Γ), and define the ε-safe homomorphism $\mu : \widehat{\Sigma}^* \to \widehat{\Delta}^*$, for $\Delta = \Gamma \cup A$, by $\alpha \mapsto g(\alpha)h(\alpha)\overline{h(\alpha)}$. For all $u \in \widehat{\Sigma}^*$, $\mu(u) \in \mathcal{D}_\Delta$ iff $u \in g^{-1}(\mathcal{D}_\Gamma)$; in addition, $\pi_A(\mu(u)) = h(u)$, with π_A being the projection of $\widehat{\Delta}^*$ into A^*. Then we have:

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

We conclude by emphasizing that $\mu(K)$ is an ε-CFL from Proposition 3.11.

($\Leftarrow$) Obvious from Theorem 5.6 and Proposition 4.5, by choosing g to be the identity mapping. □

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

5.3. Characterizations of indexed languages by rational transductions

Let us recall the characterization of context-free languages by rational transductions: there exists a language A such that for every language L, L is context-free iff $L = \tau(\mathcal{D}_A)$ for some rational transduction τ.

Theorem 5.6 extend this result to indexed language by considering ε-CFTs rather than rational transductions. We now prove another extension, by applying rational transductions to an indexed language of Dyck words that we consider to be the *indexed Dyck language*.

Let us first remark that Nivat's Theorem and Corollary 5.8 imply the following.

Proposition 5.11. *A language is indexed iff there is a rational transduction τ, an alphabetic ε-safe homomorphism g and alphabets Γ_1, Γ_2 such that*

$$L = \tau(\mathcal{D}_{\Gamma_1} \cap g^{-1}(\mathcal{D}_{\Gamma_2})).$$

We then simply have to prove that there are an alphabetic ε-safe homomorphism g and alphabets Γ_1, Γ_2 such that every indexed language can be written $L = \tau(\mathcal{D}_{\Gamma_1} \cap g^{-1}(\mathcal{D}_{\Gamma_2}))$.

Definition 5.12. Let $A_1 = \{a_1, b_1\}$, $A_2 = \{a_2, b_2\}$, $A'_1 = \{a'_1, b'_1\}$ be pairwise disjoint alphabets, and $A = A_2 \cup A_1 \cup A'_1$. Let $\sigma : \widehat{A}^* \longrightarrow \widehat{A_1}^*$ be the alphabetic ε-safe homomorphism such that

$$\sigma(\alpha) = \varepsilon \text{ if } \alpha \in A_2, \quad \sigma(\alpha) = \alpha \text{ if } \alpha \in A_1, \quad \sigma(\alpha') = \bar{\alpha} \text{ if } \alpha \in A_1,$$

we define the *indexed Dyck language*: $\mathcal{D}_{A_1,A_2} = \mathcal{D}_A \cap \sigma^{-1}(\mathcal{D}_{A_1})$.

Then, $\mathcal{D}_{A_1,A_2}$ is the set of all Dyck words such that if we inverse letters of A'_1 and remove letters of A_2, the obtained word is still a Dyck word. Note that $\mathcal{D}_{A_1,A_2}$ is an indexed language from Corollary 5.8.

Example 5.13. The word $u = b_1 a_2 b'_1 \bar{b'_1} \bar{a_2} \bar{b_1}$ belongs to $\mathcal{D}_{A_1,A_2}$ since $u \in \mathcal{D}_{A_1 \cup A_2 \cup A'_1}$ and $\sigma(u) = b_1 \bar{b_1} b_1 \bar{b_1} \in \mathcal{D}_{A_1}$.

Proposition 5.14. *For every alphabets Γ_1, Γ_2, and every alphabetic ε-safe homomorphism $g : \widehat{\Gamma_2}^* \to \widehat{\Gamma_1}^*$, there is an ε-safe homomorphism $h : \widehat{\Gamma_2}^* \to \widehat{A}^*$ such that*

$$\mathcal{D}_{\Gamma_2} \cap g^{-1}(\mathcal{D}_{\Gamma_1}) = h^{-1}(\mathcal{D}_{A_1,A_2}).$$

Proof. We suppose Γ_1 and Γ_2 to be disjoint and ordered as follows: $\Gamma_1 = \{c_1, \dots, c_n\}$ and $\Gamma_2 = \{d_1, \dots, d_m\}$ and we define the injective ε-safe homomorphism $\mu : (\widehat{\Gamma_1} \cup \widehat{\Gamma_2})^* \to (\widehat{A_1} \cup \widehat{A_2})^*$:

$$c_i \mapsto a_1 {b_1}^i a_1, \qquad d_i \mapsto a_2 {b_2}^i a_2, \qquad \text{and for all } a \in \Gamma_1 \cup \Gamma_2, \bar{a} \mapsto \overline{\mu(a)}.$$

Claim 1. $\mathcal{D}_{\Gamma_1 \cup \Gamma_2} = \mu^{-1}(\mathcal{D}_{A_1 \cup A_2})$, $\mathcal{D}_{\Gamma_1} = \mu^{-1}(\mathcal{D}_{A_1})$ *and* $\mathcal{D}_{\Gamma_2} = \mu^{-1}(\mathcal{D}_{A_2})$.

Now, we define the ε-safe homomorphism $h : \widehat{\Gamma_2}^* \to \widehat{A}^*$ as follows: for all $a \in \Gamma_2$,

$$a \mapsto \mu(a)\mu(g(a)) \text{ if } g(a) \in \Gamma_1 \cup \{\varepsilon\}, \qquad a \mapsto \mu(a)\overline{\mu(g(a))'} \text{ if } g(a) \in \overline{\Gamma_1},$$
$$\text{and } \bar{a} \mapsto \overline{h(a)}.$$

Note that in the definition above, we use the following notation: if $u = \alpha_1 \dots \alpha_n \in \widehat{A_1}^*$, then $u' = \alpha'_1 \dots \alpha'_n$.

Claim 2. $\sigma \circ h = \mu \circ g$.

Proof of Claim 2. For all $a \in \Gamma_2$, $\mu(a) \in {A_2}^*$ and then $\sigma(\mu(a)) = \varepsilon$. In addition,

- if $g(a) \in \Gamma_1 \cup \{\varepsilon\}$, then $\mu(g(a)) \in {A_1}^*$ and $\sigma(\mu(g(a)) = \mu(g(a))$. It follows that

$$\sigma(h(a)) = \sigma(\mu(a))\sigma(\mu(g(a))) = \mu(g(a));$$

- if $g(a) \in \overline{\Gamma_1}$, $\mu(g(a)) \in \overline{A_1}^*$, then $\overline{\mu(g(a))'} \in {A'_1}^*$ and $\sigma(\overline{\mu(g(a))'}) = \mu(g(a))$. It follows that

$$\sigma(h(a)) = \sigma(\mu(a))\sigma(\overline{\mu(g(a))'}) = \mu(g(a));$$

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444

- $\sigma(h(\bar{a})) = \overline{\sigma(h(a))} = \overline{\mu(g(a))} = \mu(g(\bar{a}))$. □

Claim 3. $\mathcal{D}_{\Gamma_2} = h^{-1}(\mathcal{D}_A)$.

Proof of Claim 3. Clearly, since for all $a \in \Gamma_2$, $h(a) \in A^*$ and $h(\bar{a}) = \overline{h(a)}$, for every $u \in \mathcal{D}_{\Gamma_2}$, $h(u) \in \mathcal{D}_A$.

Conversely, suppose that $h(u) \in \mathcal{D}_A$: the projection $\pi_{\widehat{A_2}}(h(u))$ belongs to $\mathcal{D}_{A_2}$. From construction of h, $\pi_{\widehat{A_2}}(h(u)) = \mu(u)$ and then from Claim 1, $u \in \mathcal{D}_{\Gamma_2}$. □

We are now ready to conclude:

$$\begin{aligned}\mathcal{D}_{\Gamma_2} \cap g^{-1}(\mathcal{D}_{\Gamma_1}) &= h^{-1}(\mathcal{D}_A) \cap g^{-1}(\mu^{-1}(\mathcal{D}_{A_1})) && \text{from Claims 3 and 1}\\ &= h^{-1}(\mathcal{D}_A) \cap h^{-1}(\sigma^{-1}(\mathcal{D}_{A_1})) && \text{from Claim 2}\\ &= h^{-1}(\mathcal{D}_A \cap \sigma^{-1}(\mathcal{D}_{A_1})) = h^{-1}(\mathcal{D}_{A_1,A_2}). && \square\end{aligned}$$

Combining Propositions 5.11 and 5.14 we get that a language L is indexed iff there is a rational transduction τ such that $L = \tau(\mathcal{D}_{A_1,A_2})$.

Theorem 5.15. *The family of indexed languages is the principal full trio generated by* $\mathcal{D}_{A_1,A_2}$.

6. Two sided Dyck context-free transductions

Let us extend the class of TCFLs to transducers:

Definition 6.1. A two-sided-Dyck context-free transduction (TCFT) is a context-free transduction whose domain is a TCFL.

Since the domain of a context-free transduction is a context-free language, a TCFT is simply a context-free transduction whose domain contains only two-sided Dyck words.

We have already seen that TCFLs are strictly more expressive than ε-CFLs, so TCFTs are strictly more expressive than ε-CFTs. We prove however that ranges of the Dyck language by TCFTs and by ε-CFLs define in fact the same classes, and then TCFTs can be used to generate indexed languages.

The reason to consider TCFTs instead of ε-CFTs is that it is decidable to know whether a context-free transduction is a TCFT, while it is undecidable to know whether it is an ε-CFT (from Theorem 3.23).

Our proof is based on the fact if a transduction grammar is reduced (see Proposition 6.2) and generates a TCFT, the ρ-reduction of the domain of the transduction derived from any nonterminal symbol is a singleton.

Proposition 6.2 *([19]). For every context-free (transduction) grammar, one can effectively construct a equivalent context-free (transduction) grammar satisfying: every nonterminal X derives at least a terminal word and is accessible by a derivation from S. A such a grammar is called* **reduced***.*

Lemma 6.3. *Let $G = (N, \widehat{\Gamma}, \Sigma, P, S)$ be a reduced context-free transduction grammar generating a TCFT. For $X \in N$, there exists $w_X \in \widehat{\Gamma}^*$ such that for all $(\omega, u) \in \widehat{\Gamma}^* \times \Sigma^*$, if $X \xrightarrow{*} (\omega, u)$, then $\rho(\omega) = w_X$.*

In addition, w_X is effectively computable.

Proof. Since G generates a TCFT, if $S \xrightarrow{*} (\omega, u)$ then $\rho(\omega) = \varepsilon$ and hence, $w_S = \varepsilon$. Suppose now that $X \xrightarrow{*} (\omega_1, u_1)$ and $X \xrightarrow{*} (\omega_2, u_2)$. Since G is reduced, there is a derivation $S \xrightarrow{*} \omega_3 X \omega_4$ and then $\rho(\omega_3\omega_1\omega_4) = \rho(\omega_3\omega_2\omega_4) = \varepsilon$. It follows that $\rho(\omega_1) = \rho(\omega_2) = \rho(\bar{\omega_3}\bar{\omega_4})$.

Using the following procedure, we compute every word w_X:

while $N \neq \emptyset$, do

1. *choose a terminal production $X \longrightarrow (\omega, u) \in P$;*
2. *state $w_X = \rho(\omega)$;*
3. *remove from P every production whose LHS is X;*
4. *replace X by (w_X, ε) in every RHS of productions in P; $N := N - \{X\}$.*

Clearly, the procedure terminates, and preserves the following invariant: for all $X \in N$, there is (ω, u) such that $X \xrightarrow{*}_P (\omega, u)$ iff $\rho(\omega) = w_X$. □

The words w_X will be the key tool of our construction. We will also need few properties of Dyck words that we state below.

Lemma 6.4. *Let us call* Dyck factor *every word $\omega \in \widehat{\Gamma}^*$ for which there exist $\omega_1, \omega_2 \in \widehat{\Gamma}^*$ such that $\omega_1\omega\omega_2 \in \mathcal{D}_\Gamma$. The following properties hold:*

1. *for all $\omega \in \widehat{\Gamma}^*$, ω is a Dyck factor iff $\rho^+(\omega) \in \overline{\Gamma}^*\Gamma^*$;*
2. *if $\omega_1\omega_2 \in \widehat{\Gamma}^*$ and $\rho^+(\omega_1\omega_2) \in \overline{\Gamma}^*\Gamma^*$, then $\rho^+(\omega_1), \rho^+(\omega_2) \in \overline{\Gamma}^*\Gamma^*$;*
3. *for all $\omega \in \widehat{\Gamma}^*$, if $\rho(\omega) = w$ and $\rho^+(\omega) \in \overline{\Gamma}^*\Gamma^*$ then $\rho^+(\omega\bar{w}w) = \rho^+(\omega)$.*

Proof. 1. If there are $x, y \in \Gamma^*$ such that $\rho^+(\omega) = \bar{x}y$, then $x\omega\bar{y} \in \mathcal{D}_\Gamma$, since $\rho^+(x\omega\bar{y}) = \rho^+(x\rho^+(\omega)\bar{y}) = \rho^+(x\bar{x}y\bar{y}) = \varepsilon$.
Conversely, suppose that $\omega_1\omega\omega_3 \in \mathcal{D}_\Gamma$:
- $\rho^+(\omega_1\omega\omega_3) = \rho^+(\rho^+(\omega_1)\rho^+(\omega\omega_3)) = \varepsilon$ and then $\rho^+(\omega_1) \in \Gamma^*$;
- $\rho^+(\omega_1\omega\omega_3) = \rho^+(\rho^+(\omega_1\omega)\rho^+(\omega_3)) = \varepsilon$ and then $\rho^+(\omega_1\omega) \in \Gamma^*$.

It follows that $\rho^+(\omega) \in \overline{\Gamma}^*\Gamma^*$.

2. From item (1) $\omega_1\omega_2$ is a Dyck factor, and then so are ω_1 and ω_2. Again from (1), we get that $\rho^+(\omega_1), \rho^+(\omega_2) \in \overline{\Gamma}^*\Gamma^*$.
3. Suppose that $\rho(\omega) = w$ and $\rho^+(\omega) \in \overline{\Gamma}^*\Gamma^*$. There are then $u_1, u_2, u_3 \in \Gamma^*$ such that $\rho^+(\omega) = \bar{u_1}\,\bar{u_2}u_2u_3$ and $\rho(\omega) = \overline{u_1}u_3$. Then $\rho^+(\omega\bar{w}w) = \rho^+(\rho^+(\omega)\bar{w}w) = \rho^+(\bar{u_1}\,\bar{u_2}u_2u_3\bar{u_3}u_1\bar{u_1}u_3) = \rho^+(\bar{u_1}\bar{u_2}u_2u_3) = \rho^+(\omega)$. □

Proposition 6.5. *For every TCFT $\tau \subseteq \widehat{\Gamma}^* \times \Sigma^*$, one can effectively find an ε-CFT τ' such that $\tau(\mathcal{D}_\Gamma) = \tau'(\mathcal{D}_\Gamma)$.*

Proof. According to Proposition 6.2, we assume τ to be generated by a reduced CFTG $G = (N, \widehat{\Gamma}, \Sigma, P, S)$. We also suppose, without loss of generality, that every production $X \longrightarrow \Omega \in P$ satisfies $\Omega \in NN \cup (\widehat{\Gamma}^* \times \Sigma^*)$. According to Lemma 6.3, for every $X \in N$, we can effectively construct the unique word w_X such that for all $(\omega, u) \in \widehat{\Gamma}^* \times \Sigma^*$, if $X \xrightarrow{*} (\omega, u)$, then $\rho(\omega) = w_X$.

We consider the context-free transduction grammar $G' = (N \cup \{N_\varepsilon\}, \widehat{\Gamma}, \Sigma, P', S)$ such that $N_\varepsilon = \{X_\varepsilon \mid X \in N\}$ and:

$$\begin{aligned} P' \;&=\; \{X \longrightarrow (\omega, u) \in P\} \cup \{X_\varepsilon \longrightarrow X(\bar{w_X}, \varepsilon) \mid X \in N\} \cup \\ &\quad\; \{X \longrightarrow Y_\varepsilon(w_Y, \varepsilon)Z_\varepsilon(w_Z, \varepsilon) \mid X \longrightarrow YZ \in P\}. \end{aligned}$$

Claim 1. *For every $X \in N$, and every $(\omega, u) \in \widehat{\Gamma}^* \times \Sigma^*$:*

1. *if $X \xrightarrow{*}_{G'} (\omega, u)$ then $\rho(\omega) = w_X$;*
2. *if $X_\varepsilon \xrightarrow{*}_{G'} (\omega, u)$ then $\rho(\omega) = \varepsilon$.*

Proof of Claim 1. By induction on the length of derivations:

- If $X \longrightarrow_{G'} (\omega, u)$, then $X \longrightarrow_G (\omega, u)$ by definition of P' and $\rho(\omega) = w_X$ by definition of w_X.
- Suppose that

$$\begin{aligned} X &\longrightarrow_{G'} Y_\varepsilon(w_Y, \varepsilon)Z_\varepsilon(w_Z, \varepsilon) \\ &\xrightarrow{*}_{G'} (\omega_1, u_1)(w_Y, \varepsilon)(\omega_2, u_2)(w_Z, \varepsilon) = (\omega, u). \end{aligned}$$

By induction hypothesis, $\rho(\omega_1) = \rho(\omega_2) = \varepsilon$, then $\rho(\omega) = \rho(w_Y w_Z)$. From definition of P', there is a production $X \longrightarrow YZ \in P$ and then $\rho(w_Y w_Z) = w_X$.
- Suppose that $X_\varepsilon \longrightarrow_{G'} X(\bar{w_X}, \varepsilon) \xrightarrow{*}_{G'} (\omega_1\bar{w_X}, u)$. By induction hypothesis $\rho(\omega_1) = \rho(w_X)$, then $\rho(\omega_1\bar{w_X}) = \varepsilon$. □

Claim 2. *For all $X \in N$, $u \in \Sigma^*$, and $\omega \in \widehat{\Gamma}^*$ such that $\rho^+(\omega) \in \overline{\Gamma}^*\Gamma^*$:*
if $X \xrightarrow{*}_G (\omega, u)$, then there exists $\omega' \in \widehat{\Gamma}^*$ such that

$$X \xrightarrow{*}_{G'} (\omega', u) \text{ and } \rho^+(\omega') = \rho^+(\omega).$$

Proof of Claim 2. We prove the claim by induction on the length of derivations. The basis case is trivially true since terminal productions have not been modified. For the induction case, there only one possible first step:
$X \longrightarrow_G YZ \xrightarrow{*}_G (\omega_1, u_1)Z \xrightarrow{*}_G (\omega_1\omega_2, u_1u_2)$ and $\rho^+(\omega_1\omega_2) \in \overline{\Gamma}^*\Gamma^*$. Then, from construction of G':

$$\begin{aligned} X &\longrightarrow_{G'} Y_\varepsilon(w_Y, \varepsilon)Z_\varepsilon(w_Z, \varepsilon) \\ &\longrightarrow_{G'} Y(\bar{w_Y}w_Y, \varepsilon)Z_\varepsilon(w_Z, \varepsilon) \\ &\longrightarrow_{G'} Y(\bar{w_Y}w_Y, \varepsilon)Z(\bar{w_Y}w_Z, \varepsilon). \end{aligned}$$

We have supposed that $\rho^+(\omega_1\omega_2) \in \overline{\Gamma}\Gamma^*$, then from Lemma 6.4(2), $\rho^+(\omega_1), \rho^+(\omega_2) \in \overline{\Gamma}^*\Gamma^*$ and we can then apply the induction hypothesis: there are $\omega_1', \omega_2' \in \widehat{\Gamma}^*$ such that $\rho^+(\omega_1') = \rho^+(\omega_1)$, $\rho^+(\omega_2') = \rho^+(\omega_2)$, $Y \xrightarrow{*}_{G'} (\omega_1', u_1)$ and $Z \xrightarrow{*}_{G'} (\omega_2', u_2)$. It follows that

$$X \xrightarrow{*}_{G'} (\omega_1' \bar{w}_Y w_Y \omega_2' \bar{w}_Y w_Z, u_1u_2) = (\omega', u).$$

From Claim 1, $\rho(\omega_1') = w_Y$ and $\rho(\omega_2') = w_Z$, then from Lemma 6.4(3), $\rho^+(\omega') = \rho^+(\omega_1'\omega_2')$ and so $\rho^+(\omega') = \rho^+(\omega_1\omega_2)$. □

Claim 3. *For all $X \in N$, $u \in \Sigma^*$, and $\omega \in \widehat{\Gamma}^*$ such that $\rho^+(\omega) \in \overline{\Gamma}^*\Gamma^*$:*
if $X \xrightarrow{*}_{G'} (\omega, u)$, then there exists $\omega' \in \widehat{\Gamma}^*$ such that

$$X \xrightarrow{*}_{G} (\omega', u) \text{ and } \rho^+(\omega') = \rho^+(\omega).$$

The proof of Claim 3 is similar to that of Claim 2 and uses the same arguments.

Claim 4. $\mathcal{T}_G(\mathfrak{D}_\Gamma) = \mathcal{T}_{G'}(\mathfrak{D}_\Gamma)$.

Proof of Claim 4. Direct from Claims 2 and 3 and the fact that for all $\omega \in \mathfrak{D}_\Gamma$, $\rho^+(\omega) = \varepsilon \in \overline{\Gamma}^*\Gamma^*$. □

We are now ready to construct the required ε-CFTG. Recall that productions of G' have the form:

$$X \longrightarrow (\omega, u),\quad X \longrightarrow Y_\varepsilon(w_Y, \varepsilon)Z_\varepsilon(w_Z, \varepsilon),\quad X_\varepsilon \longrightarrow X(\bar{w}_X, \varepsilon), \text{ for } X, Y, Z \in N.$$

We construct from G' a grammar $G'' = (N'', \widehat{\Gamma}, \Sigma, P'', S)$ as follows:

1. we replace every production $X_\varepsilon \longrightarrow X(\bar{w}_X, \varepsilon) \in P'$, $X \neq S$, by the set $\{X_\varepsilon \longrightarrow \Omega(\bar{w}_X, \varepsilon) \mid X \longrightarrow \Omega \in P'\}$;
2. we remove every production whose LHS belongs to $N - \{S\}$.

We can easily see that for all $X \in N''$, $\eta_{G''}(X) = \varepsilon$, then from Proposition 4.3, $\mathcal{T}_{G''}$ is an ε-CFT. In addition, G'' and G' generate the same transduction, then from Claim 4, $\mathcal{T}_G(\mathfrak{D}_\Gamma) = \mathcal{T}_{G''}(\mathfrak{D}_\Gamma)$. □

As direct consequences of Proposition 6.5, Theorem 4.5 and Corollary 5.9, we get the two following theorems.

Theorem 6.6. *Given a TCFT $\tau \subseteq \widehat{\Gamma}^* \times A^*$, the following assertions hold:*

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

$$\tau(\mathfrak{D}_\Gamma) = h(g^{-1}(\mathfrak{D}_\Gamma) \cap X)$$

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

$$\tau(\mathfrak{D}_\Gamma) = h(g^{-1}(\mathfrak{D}_\Gamma) \cap R \cap \mathfrak{D}_\Delta).$$

Theorem 6.7. *A language L is indexed iff there is a TCFT τ such that $L = \tau(\mathfrak{D}_2)$.*

We end the paper with a positive decidability result:

Theorem 6.8. *The problem to know whether a context-free transduction is a TCFT is decidable in polynomial time.*

Proof. Given a context-free transduction $\tau \subseteq \widehat{\Gamma}^* \times \Sigma^*$, τ is a TCFT iff $\mathrm{Dom}(\tau) \in \mathfrak{I}_\Gamma$ where $\mathrm{Dom}(\tau)$ is context-free language. This problem is decidable in polynomial time from Theorem 3.24. □

Declaration of competing interest

No competing interest.

References

[1] N. Chomsky, M.P. Schützenberger, The algebraic theory of context-free languages, in: Computer Programming and Formal Systems, Studies in Logic, North-Holland Publishing, 1963, pp. 118–161.

[2] A.N. Maslov, Hierarchy of indexed languages of arbitrary level, Sov. Math. Dokl. 115 (14) (1974) 1170–1174.

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

[4] T. Knapik, D. Niwinski, P. Urzyczyn, Higher-order pushdown trees are easy, in: FOSSACS, in: Lecture Notes in Comput. Sci., vol. 2303, Springer, 2002, pp. 205–222.

[5] M. Hague, A.S. Murawski, C.L. Ong, O. Serre, Collapsible pushdown automata and recursion schemes, in: LICS, Proceedings, IEEE Computer Society, 2008, pp. 452–461.

[6] D. Caucal, On infinite transition graphs having a decidable monadic theory, Theor. Comput. Sci. 290 (1) (2003) 79–115.

[7] L. Ong, Higher-order model checking: an overview, in: LICS, IEEE Computer Society, 2015, pp. 1–15.

[8] A.N. Maslov, Multilevel stack automata, Probl. Inf. Transm. 12 (1976) 38–43.

[9] S. Fratani, Automates à piles de piles ... de piles, Ph.D. thesis, Université Bordeaux 1, 2005.

[10] S. Hirose, M. Nasu, Left universal context-free grammars and homomorphic characterizations of languages, Inf. Control 50 (2) (1981) 110–118.

[11] J. Berstel, L. Boasson, Balanced grammars and their languages, in: Formal and Natural Computing, in: Lecture Notes in Comput. Sci., vol. 2300, Springer, 2002, pp. 3–25.

[12] D. Weir, Characterizing Mildly Context-Sensitive Grammar Formalisms, Ph.D. thesis, University of Pennsylvania, available as Technical Report MS-CIS-88-74, 1988.

[13] M. Kanazawa, Multidimensional trees and a Chomsky–Schützenberger–Weir representation theorem for simple context-free tree grammars, J. Log. Comput. 26 (5) (2016) 1469–1516.

[14] A. Sorokin, Monoid automata for displacement context-free languages, CoRR, arXiv:1403.6060.

[15] J. Duske, R. Parchmann, J. Specht, A homomorphic characterization of indexed languages, Elektron. Inf.verarb. Kybern. 15 (1979) 187–195.

[16] R. Parchmann, Balanced context-free languages and indexed languages, Elektron. Inf.verarb. Kybern. 20 (10/11) (1984) 543–556.

[17] R. Parchmann, J. Duske, Grammars, derivation modes and properties of indexed and type-0 languages, Theor. Comput. Sci. 49 (1987) 23–42.

[18] J. Duske, R. Parchmann, Linear indexed languages, Theor. Comput. Sci. 32 (1) (1984) 47–60.

[19] J. Berstel, L. Boasson, Context-free languages, in: Handbook of Theoretical Computer Science, vol. B, MIT Press, 1990, pp. 59–102.

[20] A. Okhotin, Non-erasing variants of the Chomsky-Schützenberger theorem, in: Developments in Language Theory, in: Lecture Notes in Comput. Sci., vol. 7410, Springer, 2012, pp. 121–129.

[21] A. Bertoni, C. Choffrut, R. Radicioni, The inclusion problem of context-free languages: some tractable cases, in: DLT 2009, Proceedings, in: Lecture Notes in Comput. Sci., vol. 5583, Springer, 2009, pp. 103–112.

[22] J. Hopcroft, J.D. Ullman, Introduction to Automata Theory, Languages and Computation, Addison-Wesley, 1979.

[23] J. Berstel, Transductions and Context-Free Languages, Teubner Verlag, 1979.

[24] M. Nivat, Transductions des langages de Chomsky, Ann. Inst. Fourier 18 (1) (1968) 339–455.

[25] M.J. Fischer, Grammars with macro-like productions, in: 9th Annual Symposium on Switching and Automata Theory, IEEE Computer Society, 1968, pp. 131–142.

Please cite this article in press as: S. Fratani, E.M. Voundy, Epsilon-reducible context-free languages and characterizations of indexed languages, Inf. Comput. (2019), https://doi.org/10.1016/j.ic.2019.104444