The Dissecting Power of Regular Languages

Tomoyuki Yamakami* and Yuichi Kato*

Abstract. A recent study on structural properties of regular and context-free languages has greatly promoted our basic understandings of the complex behaviors of those languages. We continue the study to examine how regular languages behave when they are asked to halve numerous infinite languages. A particular interest rests on a situation in which a regular language needs to "dissect" a given infinite language into two subsets of infinite sizes. Every context-free language is dissected by carefully chosen regular languages (or it is REG-dissectible). In a larger picture, we show that constantly-growing languages and semi-linear languages are REG-dissectible. Under certain natural conditions, complements and finite intersections of semi-linear languages also become REG-dissectible. Restricted to bounded languages, the intersections of finitely many context-free languages and, more surprisingly, the entire Boolean hierarchy over bounded context-free languages are REG-dissectible. As an immediate application of REG-dissectibility, we show another structural property, in which an appropriate bounded context-free language can "separate with infinite margins" two given nested infinite bounded context-free languages.

keywords. regular language, context-free language, bounded language, semi-linear, constant growth, dissectible, i-separate

arXiv:1202.4883v2 [cs.FL] 8 Sep 2012

1 Background Knowledge and the Results' Overview

The exquisitely complex behaviors of formal languages are often dictated by multiple-layers of inner structures of the languages and a mathematical theory over those languages has been developed in the past six decades alongside the discovery of some of the hidden structures. In an early stage of the study of context-free languages, for instance, a notion of *semi-linearity*—a structural property on the frequency of occurrences of symbols—was found in [7] and a *pumping lemma*—another property regarding the growth rate of strings—was proven in [1]. Similarly, underlying structures of regular languages have been analyzed within a number of different frameworks, including the Myhill-Nerode theorem, monadic second-order logic, and finitely generated monoids. Recently, new realms of structural properties of languages have been studied by obvious analogy with structural complexity issues of polynomial time-bounded complexity classes. Such properties include *primeimmunity* as well as *pseudorandomness* against the regular and context-free languages, introduced in [9], and a notion of *minimal cover*, which was applied to the regular languages in [2]. In the literature, numerous key questions concerning the behaviors of languages have been raised but left unsolved. We suspect that the difficulty in answering those questions may be rooted in yet-unknown structures that constitute the languages.

To promote our understandings of formal languages in general, it may be desirable to unearth the hidden structural properties of the languages. In this line of study, this paper aims at exploring another structural property, which is seemingly innocent but possibly fundamental, concerning the ability to partition a target infinite set into two portions of infinite sizes. This simple property, which we name "dissectibility," seems more suitable for weak computations, because, as shown in Section 3, polynomial-time decidable languages, for instance, are powerful enough to dissect any recursive languages of infinite sizes. Among models of weak computations, we are focused on the regular languages, because they are generally regarded as weak in recognition power; however, they could exhibit surprisingly high power in dissecting infinite languages. To be more precise at this point, an infinite set C is said to *dissect* a target infinite set L, as illustrated in Fig.1, if two disjoint sets $C \cap L$ and $\overline{C} \cap L$ $(= L - C)$ are both infinite, where $\overline{C}$ expresses the *complement* of C. When C is particularly a regular language, we succinctly say that L is *REG-dissectible*. We are mostly interested in clarifying exactly what kind of languages are REG-dissectible. A typical example of REG-dissectible language is the aforementioned context-free languages (Corollary 4.2). As for another example, let us consider a language L_1 generated by a grammar whose productions include a special form $S \to SS$, where S is the start symbol. Irrelevant to its computational complexity, the language L_1 can be dissected by a regular language composed of strings of lengths that are equal to zero modulo 3, because L_1 contains

*Present Affiliation: Department of Information Science, University of Fukui, 3-9-1 Bunkyo, Fukui 910-8507, Japan.

a series of strings of lengths $2k, 3k, 4k, \ldots$ for an appropriately chosen constant $k > 0$. A more concrete example is a language $L_2 = \{w^{n!} \mid w \in \{a, b\}^2, n \in \mathbb{N}\}$. Although this language L_2 is not even context-free, it can be easily dissected by a regular language consisting of strings, each of which begins with the letter a. The third example language is $L_3 = \{(ab^n)^n \mid n \in \mathbb{N}\}$, whose complement is context-free. This language L_3 can be easily dissected by a regular language whose strings contain an even number of a's. As a relevant notion, a $\mathcal{C}$-pseudorandom language [9] also dissects any language in $\mathcal{C}$ with quite large *margins*, where the intuitive term "margin" refers to the difference between two given sets.

Through Sections 3 to 4, two wider families of languages, *constantly-growing languages* and *semi-linear languages*, will be shown to be REG-dissectible. Under certain natural conditions, the complements, the intersections, and the differences of semi-linear languages are shown to be REG-dissectible using a simple analysis of length patterns of strings inside a given language. The analysis involves a manipulation of solutions of semi-linear equations and those conditions are indeed necessary to guarantee the REG-dissectibility. On the contrary, a rather obvious limitation exists for the REG-dissectibility; namely, as shown in Section 3, there is a logarithmic-space computable language that cannot be REG-dissectible. Taking a step further forward, when limited to *bounded languages* of Ginsburg and Spanier [4], we will be able to show that the intersections of finitely many context-free languages are dissected by appropriate regular languages, despite the fact that the intersections of k bounded context-free languages for $k \geq 1$ form an infinite hierarchy within the family of context-sensitive languages [6]. By elaborating our argument further, we will prove that the entire *Boolean hierarchy* over the class of bounded context-free languages is also REG-dissectible. These results will be presented in Section 4.

The REG-dissectibility notion has intimate connections to other notions. Earlier, Domaratzki, Shallit, and Yu [2] studied a notion of minimal cover, which means the "smallest" superset A of a given set B, where "smallest" means that there is no set between A and B with infinite margins. Motivated by their notion and results, we pay a special attention to a structural property of separating two infinite "nested" languages with infinite margins. In our term of "separation with infinite margins" (or *i-separation*, in short), we actually mean, as illustrated in Fig.2, that a pair of infinite sets A and B, denoted by $\mathrm{i}(B, A)$, for which A covers B with an infinite margin can be separated by an appropriate set C that lies in between the two sets with infinite margins. As an immediate application of the aforementioned REG-dissectibility results for the bounded context-free languages, we will show in Section 5 that two bounded context-free languages can be i-separated by bounded context-free languages. This i-separation result will be further extended into any level of the Boolean hierarchy over bounded context-free languages.

From the next section, we will formally introduce the notions of the REG-dissectibility and the *i*-separation and we will present detailed proofs of our major results mentioned above.

2 Notions and Notations

We briefly explain a set of basic notions and notations that will be used in the subsequent sections. First, we denote by $\mathbb{N}$ the set of all *natural numbers* (i.e., nonnegative integers) and we write $\mathbb{N}^+$ for $\mathbb{N} - \{0\}$. For each number $n \in \mathbb{N}^+$, the notation $[n]$ denotes the *integer interval* $\{1, 2, 3, \ldots, n\}$. Associated with three arbitrary numbers $a, b, k \in \mathbb{N}$, we define $A_{a,b,k}$ as the set $\{an + b \mid n \in \mathbb{N}, n \geq k\}$. The generic notation O denotes either an all-zero vector or an all-zero matrix of appropriate dimension. For two sets A and B, the set $\{x \in \Sigma^* \mid x \in A, x \notin B\}$ is the *difference* between A and B and is expressed as $A - B$. When A is a *countable* set, the succinct notation $|A| = \infty$ (resp., $|A| < \infty$) indicates that A is an infinite (resp., a finite) set. Given two countable sets A and B, we write $A \subseteq_{ae} B$ to mean $|A - B| < \infty$, and the notation $A =_{ae} B$ is used whenever both $A \subseteq_{ae} B$ and $B \subseteq_{ae} A$ hold.

An *alphabet* Σ is a finite nonempty set of "symbols" and a *string* over Σ is a finite sequence of symbols

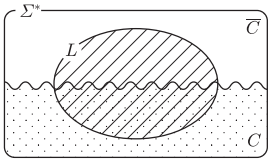


Figure 1: C dissects L.

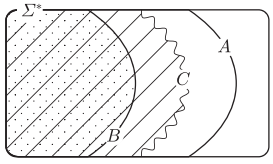


Figure 2: C i-separates $\mathrm{i}(B, A)$.

in Σ. The set of all strings over Σ is denoted Σ^* and Σ^+ expresses the set $\Sigma^* - \{\lambda\}$, where λ is the *empty string*. The *length* $|x|$ of a string x is the number of occurrences of symbols in x. For any string x and any symbol σ, the notation $\#_\sigma(x)$ stands for the number of occurrences of σ in x. Given a language S, the *length set* of S, denoted $LT(S)$, is the collection of all lengths $|x|$ for any strings x in S. We often identify a language S with its *characteristic function*, which is also denoted S (i.e., $S(x) = 1$ if $x \in S$, and $S(x) = 0$ otherwise). The sets of all regular languages and of all context-free languages are expressed respectively as REG and CFL.

The *complement* of B is the set $\Sigma^* - B$ and it is denoted $\overline{B}$ as far as its underlying alphabet Σ is clear from the context. A language is *co-infinite* if its complement is infinite. For ease of our notations, we use the following four class operations: (1) $\mathcal{C} \wedge \mathcal{D} = \{C \cap D \mid C \in \mathcal{C}, D \in \mathcal{D}\}$, (2) $\mathcal{C} \vee \mathcal{D} = \{C \cup D \mid C \in \mathcal{C}, D \in \mathcal{D}\}$, (3) $\mathcal{C} - \mathcal{D} = \{C - D \mid A \in \mathcal{C}, D \in \mathcal{D}\}$, and (4) co-$\mathcal{C} = \{\overline{C} \mid C \in \mathcal{C}\}$, where $\mathcal{C}$ and $\mathcal{D}$ are language families. For any family $\mathcal{F}$ of languages, a language S is said to be $\mathcal{F}$-*immune* if S is infinite and S has no infinite subset belonging to $\mathcal{F}$ (see, e.g., [9]).

3 How to Dissect Languages

Let us recall from Section 1 that an infinite language S is *REG-dissectible* exactly when there exists a regular language C that dissects S (i.e., $|C \cap S| = |\overline{C} \cap S| = \infty$). Moreover, a nonempty language family $\mathcal{F}$ is *REG-dissectible* if and only if every infinite language in $\mathcal{F}$ is REG-dissectible. Notice that, since this definition disregards all *finite* languages inside $\mathcal{F}$, we implicitly assume that $\mathcal{F}$ contains infinite languages. We can naturally expand the REG-dissectibility to a more general notion of $\mathcal{C}$-dissectibility simply by replacing REG with an arbitrary nonempty language family $\mathcal{C}$; however, the choice of REG is actually of great importance. In fact, it is more interesting to consider low-complexity language families like REG as a candidate for $\mathcal{C}$. One reason is that polynomial-time decidable languages, for instance, are powerful enough to dissect any infinite recursive languages.

Example 3.1 We claim that every infinite recursive language is P-dissectible, where P is the family of all polynomial-time decidable languages. Let L be any infinite language, over an alphabet Σ, recognized by a single-tape deterministic Turing machine M that eventually halts on all inputs. For simplicity, let $\Sigma = \{0, 1\}$ and assume that $L \neq_{ae} \Sigma^*$ because, otherwise, a regular set $C = \{0x \mid x \in \Sigma^*\}$ easily dissects L. Now, we define C as follows. Let $z_0, z_1, z_2, \ldots$ be a standard lexicographic order of all strings over Σ. For each string x, we go through the following procedure from round 0 to round $|x|$. Initially, we set $A = R = \emptyset$. At round i, we first recover the value $C(z_i)$ by simulating the whole process of determining $C(z_i)$. We then simulate M on the input z_i within $|x|$ steps. When $M(z_i) = 1$, we update A to $A \cup \{i\}$ if $C(z_i) = 1$, and R to $R \cup \{i\}$ if $C(z_i) = 0$. On the contrary, when either $M(z_i) = 0$ or $M(z_i)$ is not obtained within $|x|$ steps, we do nothing. After round $|x|$, if $|A| > |R|$, then define $C(x) = 0$; otherwise, define $C(x) = 1$. Clearly, C is in P. By a diagonalization argument, we can show that $|C \cap L| = |\overline{C} \cap L| = \infty$. Therefore, every infinite recursive language can be dissected by an appropriate language in P.

In the following second example, we will show that a simple use of *advice* makes it possible to dissect an arbitrary language by regular languages. For basic properties of the advice, the reader may refer to [8].

Example 3.2 We claim that every infinite language is REG/n-dissectible, where REG/n is the collection of *advised regular languages*, each of which is of the form $\{x \mid M \text{ accepts } [{}^{x}_{h(|x|)}]\}$ for an appropriate deterministic finite automaton (or a dfa), an advice alphabet Γ, and an advice function $h : \mathbb{N} \to \Gamma^*$ satisfying $|h(n)| = n$ for all $n \in \mathbb{N}$, where $[{}^{x}_{y}]$ is a *track notation* used in [8]. To verify this claim, take any infinite language L over an alphabet Σ. Since L is infinite, the length set $LT(L)$ is also infinite. Hence, we partition $LT(L)$ into two infinite subsets, say, S_1 and S_2; that is, $S_1 \cap S_2 = \emptyset$, $LT(L) = S_1 \cup S_2$, and $|S_1| = |S_2| = \infty$. We assume that $0 \notin S_1$. Now, let us define an advice function $h : \mathbb{N} \to \{0, 1\}^*$ as $h(n) = 10^{n-1}$ if $n \in S_1$ and $h(n) = 0^n$ otherwise. We also define a dfa M as follows: on input $[{}^{x}_{y}]$, if $y = h(|x|)$, then M accepts the input; otherwise, it rejects the input. The language $C = \{x \mid M \text{ accepts } [{}^{x}_{h(|x|)}]\}$ then belongs to REG/n. Obviously, for any string $x \in L$ with $|x| \in S_1$, since $h(|x|) = 10^{|x|-1}$, M accepts $[{}^{x}_{h(|x|)}]$. It thus holds that $|C \cap L| = \infty$. Similarly, for any $x \in S$ with $|x| \in S_2$, M rejects $[{}^{x}_{h(|x|)}]$, implying $|\overline{C} \cap L| = \infty$. In conclusion, C dissects L.

As noted in Section 1, a pattern of the lengths of strings in a target language surely plays a key role in proving its REG-dissectibility. This fact turns our attention to languages composed of strings satisfying a

certain length condition, known as a "constant growth property." Formally, a nonempty language L is said to be *constantly growing* if there exists a constant $p > 0$ and a finite subset $K \subseteq \mathbb{N}^+$ that meet the following condition: for every string $x \in L$ with $|x| \geq p$, there exist a string $y \in L$ and a constant $c \in K$ for which $|x| = |y| + c$ holds. Such languages can be easily dissected by appropriately chosen regular languages as shown in the next lemma.

Lemma 3.3 *Every infinite constantly-growing language is* REG*-dissectible.*

Proof. Let L be any infinite language over an alphabet Σ and assume that L is constantly growing with a constant $p > 0$ and a finite set $K \subseteq \mathbb{N}^+$. Now, let c denote the maximal element in K and set $c' = c + 1$. For each index $i \in [c]$, we take a special language $L_i = \{x \in L \mid |x| \equiv i \pmod{c'}\}$, and we wish to prove that at least two distinct indices $i_1, i_2 \in [c]$ satisfy that $|L_{i_1}| = |L_{i_2}| = \infty$. Toward a contradiction, we assume otherwise. Since $L = \bigcup_{i \in [c]} L_i$, exactly one index $i \in [c]$ makes L_i infinite. Let us fix such an index, say, i. Given any index $j \in [c]$, we define $S_{i,j} = \{y \in L \mid \exists x \in L_i \, [\, |x| = |y| + j \,]\}$. Since L is constantly growing, a set $S_{i,j}$ must be infinite for a certain index j. Note that $S_{i,j} \subseteq L_\ell$ holds for $\ell = i - j \bmod c'$. This containment implies that L_ℓ is infinite, contradicting the uniqueness of i since $i \neq \ell$. Therefore, we can choose two distinct indices $i_1, i_2 \in [c]$ for which $|L_{i_1}| = |L_{i_2}| = \infty$. Finally, we define $C = \{x \in \Sigma^* \mid |x| \equiv i_1 \pmod{c'}\}$, which is clearly regular. Since $L_{i_1} \subseteq C$ and $L_{i_2} \subseteq \overline{C}$, it obviously follows that $|C \cap L| = |\overline{C} \cap L| = \infty$. In other words, C dissects L, as requested. □

For a wider application of Lemma 3.3, it is therefore desirable to strengthen the lemma slightly. In what follows, we succinctly write CGL for the family of all constantly-growing languages and use the notion of *CGL-immunity* to describe our proposition.

Proposition 3.4 *Every language that is not* CGL*-immune is* REG*-dissectible.*

The above proposition comes from Lemma 3.3 as well as the following *transitive closure property* of REG-dissectibility: for any two infinite languages A and B, if A is REG-dissectible and $A \subseteq B$, then B is also REG-dissectible.

Luckily, a length pattern of strings in a language is not the only feature used to dissect the target language. For example, the languages L_2 and L_3 exemplified in Section 1 are not constantly growing; however, they are dissected by regular languages. Before presenting more examples of REG-dissectible languages in the next section, we will show a plausible limitation of the dissecting power of the regular languages. Following a standard convention, the notation L stands for the family of all languages that can be recognized by two-way deterministic Turing machines using a read-only input tape together with a constant number of logarithmic space-bounded read/write work tapes. In the next proposition, we will show that L contains a language that cannot be dissected by any regular languages.

Theorem 3.5 *The language family* L *is not* REG*-dissectible.*

Proof. Let us consider the unary language $S = \{0^{n!} \mid n \in \mathbb{N}\}$ over the alphabet $\Sigma = \{0\}$. Firstly, we will show that S is in L. For this purpose, it suffices to design a logarithmic-space Turing machine that recognizes S. On input of the form 0^m, the desired machine M writes m in binary on its 1st work tape using $O(\log m)$ cells and 1 on its 2nd work tape. At each round, M reads out a number, say, n in binary written on the 2nd tape and checks if m is a multiple of n using the 3rd work tape as a counter up to n. If not, then M immediately rejects the input; otherwise, it increases n by one (in binary) before entering the next round. If the machine does not rejects until n reaches m, it accepts the input.

Secondly, we want to show that no regular language can dissect S. Assume otherwise; that is, there exists an infinite language $C \in$ REG over Σ dissects S. We need the following technical property (Claim 1) of the unary regular language C on its length set $LT(C)$. Let us recall the notation $A_{a,b,k}$ and, in addition, set $\mathcal{G} = \{(a, b, k) \mid a, b, k \in \mathbb{N}, b < a\}$ for the description of the property.

Claim 1 *For any unary language C, C is regular iff there exists a finite set $G \subseteq \mathcal{G}$ for which $LT(C) = \bigcup_{(a,b,k) \in G} A_{a,b,k}$.*

Claim 1 is attributed to Parikh [7] and it guarantees the existence of a finite set G that characterizes C. Without loss of generality, we assume that $b < a$ for any $(a, b, k) \in G$.

Since $|C \cap S| = \infty$, there exists a triplet (a, b, k) in G satisfying $|\{m \mid \exists n \geq k \, [\, m! = an + b \,]\}| = \infty$. Now, we will show that $b = 0$. First, take two integers m, n with $n \geq k$ and $m > a$ satisfying $an + b = m!$.

Since $a < m$, $m! \equiv 0 \pmod{a}$ holds. From $an + b \equiv b \pmod{a}$, we obtain $b \equiv 0 \pmod{a}$. Since $b < a$, b must be zero, as requested. Moreover, it holds that $a > 1$. To see this, suppose that $a = 1$. Since $A_{1,0,k}$ equals $\{n \mid n \geq k\}$, we conclude that $|\overline{A_{1,0,k}}| < \infty$. Therefore, it follows that $|LT(\overline{C}) \cap LT(S)| < \infty$, contradicting $|\overline{C} \cap S| = \infty$.

Since $a > 1$ and $b = 0$, for a certain large constant k', it holds that $\{m! \mid m \geq k'\} \subseteq A_{a,0,k}$. This implies that $|LT(\overline{C}) \cap LT(S)| < \infty$. This is a clear contradiction, and therefore C cannot dissect S. □

For convenience, we denote by REG-DISSECT the collection of all *infinite* REG-dissectible languages. It is not difficult to prove the following closure/non-closure properties. (1) The set REG-DISSECT is closed under concatenation, reversal, Kleene star, and union. (2) REG-DISSECT is not closed under intersection with regular languages. (3) Moreover, REG-DISSECT is not closed under λ-free homomorphism as well as under quotient with regular languages, where λ is the empty string. The last two properties can be proven using certain languages derived from the one presented in the proof of Theorem 3.5.

4 Context-Free Languages and Bounded Languages

Parikh [7] discovered that the number of times that each symbol occurs in each string of a given context-free language L must satisfy certain linear Diophantine equations. This result inspired a notion of *semi-linear languages*. Context-free languages are an important example of semi-linear languages and a semi-linear nature of languages will be exploited in certain cases of the REG-dissectibility proofs of the languages. First, we will explain the notion of semi-linear sets and languages using a matrix formalism. A subset A of $\mathbb{N}^k$ is called *linear* if there exist a number $m \in \mathbb{N}$ and an $(m+1) \times k$ nonnegative integer matrix (called a *critical matrix*) T satisfying: for every point $v \in \mathbb{N}^k$, $v \in A$ iff an equation $(1, z_1, z_2, \ldots, z_m)T = v$ holds for a certain tuple (called a *solution*) $(z_1, z_2, \ldots, z_m) \in \mathbb{N}^m$. A *semi-linear set* is a union of finitely many linear sets. For any string x over an alphabet $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_k\}$, a *Parikh image* of x, denoted by $\Psi(x)$, is a point $(\#_{\sigma_1}(x), \#_{\sigma_2}(x), \ldots, \#_{\sigma_k}(x))$ in the space $\mathbb{N}^k$, and the *commutative image* (or the *Parikh image*) $\Psi(L)$ of a language L over Σ refers to the set $\{\Psi(x) \mid x \in L\}$. The language L is called *semi-linear* whenever $\Psi(L)$ is semi-linear.

The family of all semi-linear languages is denoted by SEMILIN, and SEMILIN(2) expresses the family SEMILIN $\wedge$ SEMILIN.

Lemma 4.1 SEMILIN $\subseteq$ REG-DISSECT *but* SEMILIN(2) $\nsubseteq$ REG-DISSECT.

Proof. Every semi-linear language L is defined by a set of certain linear equations and this fact proves that L has the property of constant growth. Lemma 3.3 therefore leads to the first part of the lemma. To see that SEMILIN(2) is not REG-dissectible, let us consider two example languages $L_1 = \{0^n 1^n \mid n \in \mathbb{N}\}$ and $L_2 = \{1^n 0^n \mid n \in \mathbb{N}\} \cup \{0^{n!} 1^{n!} \mid n \in \mathbb{N}\}$. Since $\Psi(L_1) = \Psi(L_2) = \{(n, n) \mid n \in \mathbb{N}\}$, L_1 and L_2 are semi-linear. However, the intersection $L_1 \cap L_2 \in$ SEMILIN, which equals $\{0^{n!} 1^{n!} \mid n \in \mathbb{N} \in$ SEMILIN$\}$, can be shown to be non-REG-dissectible by an argument similar to the proof of Theorem 3.5. □

Since CFL $\subseteq$ SEMILIN, Lemma 4.1 immediately yields the following consequence.

Corollary 4.2 *The language family* CFL *is* REG-*dissectible.*

To utilize well-studied properties on semi-linear languages, we will limit our attention within a restricted part of context-free languages. A language L over an alphabet Σ is said to be *bounded* if there are fixed nonempty strings $w_1, w_2, \ldots, w_m$ in Σ^* such that L is a subset of $L[w_1, w_2, \ldots, w_m] =_{\text{def}} \{w_1^{i_1} w_2^{i_2} \cdots w_m^{i_m} \mid i_1, i_2, \ldots, i_m \in \mathbb{N}\}$ [4]. For readability, we abbreviate as BCFL the family of all bounded context-free languages. The *k-conjunctive closure* of BCFL, denoted BCFL(k), is defined inductively as follows: BCFL(1) = BCFL and BCFL(k) = BCFL($k-1$) $\wedge$ BCFL for every index $k \geq 2$. Earlier, Liu and Weiner [6] proved that the collection $\{\text{BCFL}(k) \mid k \in \mathbb{N}^+\}$ forms an infinite hierarchy within the family of context-sensitive languages.

Theorem 4.3 *For any index $k \geq 1$,* BCFL(k) *is* REG-*dissectible.*

For the proof of Theorem 4.3, we define $\tilde{\Psi}(w)$ to be $\{(i_1, i_2, \ldots, i_m) \in \mathbb{N}^m \mid w = w_1^{i_1} w_2^{i_2} \cdots w_m^{i_m}\}$ for each string w in $L[w_1, w_2, \ldots, w_m]$. Notice that $\tilde{\Psi}(w)$ could contain numerous elements because w may have more than one expression of the form $w_1^{i_1} w_2^{i_2} \cdots w_m^{i_m}$. Finally, we define $\tilde{\Psi}(L) = \bigcup_{w \in L} \tilde{\Psi}(w)$ for any

bounded language L. This operator $\tilde{\Psi}$ works similarly as Ψ does and, by exploiting this similarity, Ginsburg [3] exhibited a close relationship between bounded context-free languages L and the semi-linearity of $\tilde{\Psi}(L)$. What we need for our proof given below is the following slightly weaker form of [3, Theorem 5.4.2]: for any subset L of $L[w_1, \ldots, w_k]$ in $L \in \mathrm{BCFL}$, $\tilde{\Psi}(L)$ is semi-linear, and thus L is in SEMILIN.

Proof of Theorem 4.3. We start with the following general claim regarding Ψ. By viewing $w_1, w_2, \ldots, w_m$ as "different" symbols $\sigma_1, \sigma_2, \ldots, \sigma_m$ as in [3], a similarity between $\Psi(w)$ and $\tilde{\Psi}(w)$ makes the claim true for $\tilde{\Psi}$ as well.

Claim 2 *For any $L_1, L_2 \in$ SEMILIN, if $|L_1 \cap L_2| = \infty$ and $\Psi(L_1) \cap \Psi(L_2) \subseteq \Psi(L_1 \cap L_2)$ hold, then $L_1 \cap L_2$ is* REG*-dissectible. More generally, let k be any number ≥ 2 and let $L_1, L_2, \ldots, L_k$ be k semi-linear languages. If $\left|\bigcap_{i=1}^{k} L_i\right| = \infty$ and $\bigcap_{i=1}^{k} \Psi(L_i) \subseteq \Psi(\bigcap_{i=1}^{k} L_i)$ hold, then $\bigcap_{i=1}^{k} L_i$ is* REG*-dissectible.*

Proof. Since $\Psi(L_1 \cap L_2) \subseteq \Psi(L_1) \cap \Psi(L_2)$ always holds, our assumption actually means $\Psi(L_1 \cap L_2) = \Psi(L_1) \cap \Psi(L_2)$. Since the set of all semi-linear sets is closed under Boolean operations (as well as projections) [5], we conclude that $L_1 \cap L_2$ belongs to SEMILIN. Lemma 4.1 implies that $L_1 \cap L_2 \in$ REG-DISSECT. The above proof can be easily extended to the case of the intersection $\bigcap_{i=1}^{k} \Psi(L_i)$ of k commutative images. □

Now, let $L' = L[w_1, w_2, \ldots, w_m]$ and take any k subsets $L_1, L_2, \ldots, L_k \in$ BCFL of L'. As noted earlier, it follows that $L_1, L_2, \ldots, L_k \in$ SEMILIN. Here, we assume that $L = \bigcap_{i=1}^{k} L_i$ is an infinite set. By Claim 2, we only need to prove that $\bigcap_{i=1}^{k} \tilde{\Psi}(L_i) \subseteq \tilde{\Psi}(\bigcap_{i=1}^{k} L_i)$. Firstly, let $v \in \bigcap_{i=1}^{k} \tilde{\Psi}(L_i)$ and fix $i \in [k]$ arbitrarily. Since the inverse image $\tilde{\Psi}^{-1}(v) = \{w \in L' \mid v \in \tilde{\Psi}(w)\}$ must be a singleton, there exists a *unique* string $w \in L'$ for which $\tilde{\Psi}^{-1}(v) = \{w\}$. From $v \in \tilde{\Psi}(L') \subseteq \tilde{\Psi}(L_i)$, we obtain $w \in L_i$. Moreover, since i is arbitrary, we conclude that w is in $\bigcap_{i=1}^{k} L_i$. It thus follows that $v \in \tilde{\Psi}(w) \subseteq \tilde{\Psi}(\bigcap_{i=1}^{k} L_i)$. In conclusion, L should be REG-dissectible. □

Without the condition $\Psi(L_1) \cap \Psi(L_2) \subseteq \Psi(L_1 \cap L_2)$ of Claim 2, nevertheless, it is impossible to prove the intersection of two semi-linear languages to be REG-dissectible since SEMILIN(2) $\nsubseteq$ REG-DISSECT.

Next, we will show the REG-dissectibility of the Boolean hierarchy over BCFL, where the *Boolean hierarchy over* BCFL is defined as follows: $\mathrm{BCFL}_1 = \mathrm{BCFL}$, $\mathrm{BCFL}_{2k} = \mathrm{BCFL}_{2k-1} \wedge$ co-BCFL, and $\mathrm{BCFL}_{2k+1} = \mathrm{BCFL}_{2k} \vee \mathrm{BCFL}$ for every number $k \in \mathbb{N}^+$. Finally, we set $\mathrm{BCFL}_{\mathrm{BH}} = \bigcup_{k \geq 1} \mathrm{BCFL}_k$.

Theorem 4.4 *The Boolean hierarchy* $\mathrm{BCFL}_{\mathrm{BH}}$ *is* REG*-dissectible.*

Proof. Since $\mathrm{BCFL}_{2k-1} \subseteq \mathrm{BCFL}_{2k}$ holds for every number $k \in \mathbb{N}^+$, it is sufficient to prove that BCFL_{2k} is REG-dissectible for all $k \in \mathbb{N}$. We will show this claim by induction on k. For the basis case of BCFL_2 $(= \mathrm{BCFL} - \mathrm{BCFL})$, let $L_1, L_2 \in$ BCFL be languages over an alphabet Σ and concentrate on the difference $L_1 - L_2$. First, we intend to prove Claim 3. In the claim, the notation $\|v\|_1$ for any vector v in a Euclidean space denotes the ℓ_1*-norm* of v; that is, $\|v\|_1 = \sum_i |v_i|$ if $v = (v_i)_i$.

Claim 3 *Let L_1 and L_2 be any two infinite semi-linear languages satisfying $\Psi(L_1) \not\subseteq_{ae} \Psi(L_2)$. If $\Psi(L_1) - \Psi(L_2) \subseteq \Psi(L_1 - L_2)$ holds, then the difference $L_1 - L_2$ is* REG*-dissectible.*

Proof. Since $\Psi(L_1)$ and $\Psi(L_2)$ are both semi-linear, the difference $\Psi(L_1) - \Psi(L_2)$ is semi-linear as well [5]. By our assumption follows the equality $\Psi(L_1) - \Psi(L_2) = \Psi(L_1 - L_2)$. There exists a series of critical matrices that characterizes $\Psi(L_1 - L_2)$. Here, we want to fix one of them, say, $T = (v_j)_{1 \leq j \leq m}$, where each v_j is a column vector. For simplicity, we assume that $v_1 \neq O$. Given each index $i \in \{0, 1\}$, let us consider a set $A_i = \{w \in \Sigma^* \mid \exists z_1 \in \mathbb{N}\,[\,(1, 2z_1 + i, 0, \ldots, 0)T = \Psi(w)\,]\}$. Since $\Psi(A_0 \cup A_1) \subseteq \Psi(L_1 - L_2)$, we conclude that $A_0 \cup A_1 \subseteq L_1 - L_2$. It is clear that A_i is infinite and the language $C_i = \{w \in \Sigma^* \mid |w| = \|(1, 2z_1 + i, 0, \ldots, 0)T\|_1\}$ is also infinite because of $A_i \subseteq C_i$. In addition, C_i is regular because every string w in C_i satisfies $|w| = \|v_0\|_1 + (2z_1 + i)\|v_1\|_1$ and it is easy to determine whether or not this is true for a given string w by running an appropriate dfa. Since $C_0 \cap C_1 = \emptyset$ and $A_i \subseteq C_i \cap (L_1 - L_2)$ for each index $i \in \{0, 1\}$, C_i must dissect $L_1 - L_2$. Hence, $L_1 - L_2$ is REG-dissectible. □

Now, we wish to claim that $\tilde{\Psi}(L_1) - \tilde{\Psi}(L_2) \subseteq \tilde{\Psi}(L_1 - L_2)$ for two arbitrary languages $L_1, L_2 \in$ BCFL. To prove this claim, take any $v \in \tilde{\Psi}(L_1) - \tilde{\Psi}(L_2)$. Since $v \in \tilde{\Psi}(L_1)$, there exists a string $w \in L_1$ for which $v \in \tilde{\Psi}(w)$. Note that $w \notin L_2$ because, otherwise, we obtain $v \in \tilde{\Psi}(w) \subseteq \tilde{\Psi}(L_2)$, a contradiction. Since $w \in L_1 - L_2$, it follows that $v \in \tilde{\Psi}(w) \subseteq \tilde{\Psi}(L_1 - L_2)$. Using a similarity between $\Psi(w)$ and $\tilde{\Psi}(w)$ as in the

proof of Theorem 4.3, we can apply Claim 3 and then obtain the REG-dissectibility of $L_1 - L_2$.

The remaining task is to deal with the induction case of BCFL_{2k} for any number $k \geq 2$. For this purpose, we will present a simple fact on the even levels of the Boolean hierarchy over BCFL.

Claim 4 *For every number $k \geq 2$, $\mathrm{BCFL}_{2k} = \mathrm{BCFL}_{2k-2} \vee \mathrm{BCFL}_2$.*

Proof. Here, we want to claim that (*) for every number $k \geq 2$, $\mathrm{BCFL}_{2k-2} \wedge \text{co-BCFL} = \mathrm{BCFL}_{2k-2}$. Write $\mathcal{F}$ for $\mathrm{BCFL}_{2k-2} \wedge \text{co-BCFL}$ for simplicity. Since $\mathrm{BCFL}_{2k-2} = \mathrm{BCFL}_{2k-3} \wedge \text{co-BCFL}$ by the definition, $\mathcal{F}$ equals $\mathrm{BCFL}_{2k-3} \wedge (\text{co-BCFL} \wedge \text{co-BCFL})$, which is actually $\mathrm{BCFL}_{2k-3} \wedge \text{co-}(\mathrm{BCFL} \vee \mathrm{BCFL})$. Since BCFL is closed under union (i.e., $\mathrm{BCFL} \vee \mathrm{BCFL} = \mathrm{BCFL}$), it follows that $\mathcal{F} = \mathrm{BCFL}_{2k-3} \wedge \text{co-BCFL}$. By the definition again, the right-hand side of this equation coincides with BCFL_{2k-2}. Therefore, Statement (*) should hold.

Recall that BCFL_{2k} equals $\mathrm{BCFL}_{2k-1} \wedge \text{co-BCFL}$, which also coincides with $(\mathrm{BCFL}_{2k-2} \vee \mathrm{BCFL}) \wedge \text{co-BCFL}$. By DeMorgan's law, it holds that $\mathrm{BCFL}_{2k} = (\mathrm{BCFL}_{2k-2} \wedge \text{co-BCFL}) \vee (\mathrm{BCFL} \wedge \text{co-BCFL})$. Statement (*) then leads to $\mathrm{BCFL}_{2k} = \mathrm{BCFL}_{2k-2} \vee \mathrm{BCFL}_2$. □

Notice that the induction hypothesis ensures the REG-dissectibility of BCFL_{2k-2}. Since BCFL_2 has been already proven to be REG-dissectible, $\mathrm{BCFL}_{2k-2} \vee \mathrm{BCFL}_2$ must be REG-dissectible by the closure property of REG-DISSECT in Section 3. By Claim 4, this family is exactly BCFL_{2k}. This completes the proof of Theorem 4.4 □

5 Separation with Infinite Margins

In this final section, we will seek a meaningful application of our previous results regarding the REG-dissectibility of certain bounded languages. To describe this application, we need to introduce extra terminology. Given two infinite sets A and B, we say that A *covers B with an infinite margin* (A *i-covers* B, or A is an *i-cover* of B, in short) if both $B \subseteq A$ and $A \neq_{ae} B$ hold. When A i-covers B, we briefly write $\mathrm{i}(B, A)$ and call it an *i-covering pair*. A language C is said to *separate* $\mathrm{i}(B, A)$ *with infinite margins* (or *i-separate* $\mathrm{i}(B, A)$, in short) if (i) $B \subseteq C \subseteq A$, (ii) $A \neq_{ae} C$, and (iii) $B \neq_{ae} C$. For convenience, we use the notation $\mathrm{i}(\mathcal{B}, \mathcal{A})$ for two language families $\mathcal{A}$ and $\mathcal{B}$ to denote the set of all i-covering pairs $\mathrm{i}(B, A)$ satisfying $A \in \mathcal{A}$ and $B \in \mathcal{B}$. Another language family $\mathcal{C}$ is said to *i-separate* $\mathrm{i}(\mathcal{B}, \mathcal{A})$ if, for every pair $\mathrm{i}(B, A) \in \mathrm{i}(\mathcal{B}, \mathcal{A})$, there exists a set in $\mathcal{C}$ that i-separates $\mathrm{i}(B, A)$.

The following is a key lemma that bridges between REG-dissectibility and i-separation.

Lemma 5.1 *Let $\mathcal{A}, \mathcal{B}$ be any two language families. Assume that $\mathcal{A} - \mathcal{B}$ is* REG*-dissectible. It then holds that, for any $A \in \mathcal{A}$ and any $B \in \mathcal{B}$, if A i-covers B, then there exists a language in $\mathcal{E}$ that i-separates $\mathrm{i}(B, A)$, where $\mathcal{E}$ expresses the set $\{B \cup (A \cap C) \mid A \in \mathcal{A}, B \in \mathcal{B}, C \in \mathrm{REG}\}$. In other words, $\mathcal{E}$ i-separates $\mathrm{i}(\mathcal{B}, \mathcal{A})$.*

Proof. Let $A \in \mathcal{A}$ and $B \in \mathcal{B}$ be two infinite languages. Let $D = A - B$ and assume that D is infinite. Our assumption guarantees the existence of a language $C \in \mathrm{REG}$ for which C dissects D. For convenience, we define $E = B \cup (A \cap C)$. Since C dissects D, it follows that $|(A \cap C) - B| = \infty$ and $|(A \cap \overline{C}) - B| = \infty$. These conditions imply that $B \subseteq E \subseteq A$ and $|A - E| = |E - B| = \infty$. Thus, E i-separates $\mathrm{i}(B, A)$. Since $C \in \mathrm{REG}$, E clearly belongs to the language family $\mathcal{E}$. □

Concerning bounded context-free languages, we can show the following i-separation result.

Theorem 5.2 *For any index $k \in \mathbb{N}^+$, BCFL_k i-separates $\mathrm{i}(\mathrm{BCFL}_k, \mathrm{BCFL}_k)$.*

Proof. Hereafter, we want to show that $\mathrm{BCFL}_k - \mathrm{BCFL}_k$ is REG-dissectible because an application of Lemma 5.1 immediately leads to the theorem. For our purpose, it suffices to prove that $\mathrm{BCFL}_k - \mathrm{BCFL}_k$ is included in $\mathrm{BCFL}_{\mathrm{BH}}$, because $\mathrm{BCFL}_{\mathrm{BH}}$ is REG-dissectible by Theorem 4.3. More strongly, we intend to demonstrate that, for any two indices $i, j \geq 1$, $\mathrm{BCFL}_i - \mathrm{BCFL}_j \subseteq \mathrm{BCFL}_{\mathrm{BH}}$.

Given an index pair $(i, j) \in \mathbb{N}^+ \times \mathbb{N}^+$, let $\mathcal{F}_{i,j} = \mathrm{BCFL}_i - \mathrm{BCFL}_j = \mathrm{BCFL}_i \wedge \text{co-BCFL}_j$ and $\mathcal{G}_{i,j} = \mathrm{BCFL}_i \wedge \mathrm{BCFL}_j$ for simplicity. We will show that $\mathcal{F}_{i,j} \subseteq \mathrm{BCFL}_{\mathrm{BH}}$ by induction on (i, j). For the basis case $(1, 1)$, since $\mathcal{F}_{1,1} = \mathrm{BCFL}_2$ holds, clearly $\mathcal{F}_{1,1}$ is a subset of $\mathrm{BCFL}_{\mathrm{BH}}$. For the next case $(2, 1)$, we first note that $\mathrm{BCFL}_4 = (\mathrm{BCFL}_2 \wedge \text{co-BCFL}_2) \vee (\mathrm{BCFL}_2 \wedge \mathrm{BCFL}_2) = \mathcal{F}_{2,1} \vee \mathcal{G}_{2,2}$. We thus obtain $\mathcal{F}_{2,1} \subseteq \mathrm{BCFL}_4$

as well as $\mathcal{G}_{2,2} \subseteq \mathrm{BCFL}_4$. For the induction case (i,j), it is enough to consider the case where $i = 2k$ and $j = 2m+1$. Similar to Claim 4, we can prove the next useful relation.

Claim 5 co-$\mathrm{BCFL}_{2k+1} = \mathrm{BCFL}_{2k-1} \vee \mathrm{BCFL}_2$.

By Claims 4 and 5, $\mathcal{F}_{2k,2m+1}$ equals $(\mathrm{BCFL}_{2k-2} \vee \mathrm{BCFL}_2) \wedge (\text{co-}\mathrm{BCFL}_{2m-1} \vee \mathrm{BCFL}_2)$, which can be transformed into $\mathcal{F}_{2k-2,2m-1} \vee \mathcal{F}_{2,2m-1} \vee \mathcal{G}_{2k-2,2} \vee \mathcal{G}_{2,2}$. By the induction hypothesis, there are two indices ℓ_1, ℓ_2 such that $\mathcal{F}_{2k-2,2m-1} \subseteq \mathrm{BCFL}_{2\ell_1}$ and $\mathcal{F}_{2,2m-1} \subseteq \mathrm{BCFL}_{2\ell_2}$. By applying Claim 4 repeatedly, we then obtain $\mathrm{BCFL}_{2\ell_1} = \bigvee_{i=1}^{\ell_1} \mathrm{BCFL}_2$ and $\mathrm{BCFL}_{2\ell_2} = \bigvee_{i=1}^{\ell_2} \mathrm{BCFL}_2$. Likewise, we obtain $\mathrm{BCFL}_{2k-2} = \bigvee_{i=1}^{k-1} \mathrm{BCFL}_2$. Hence, $\mathcal{G}_{2k-2,2}$ equals $(\bigvee_{i=1}^{k-1} \mathrm{BCFL}_2) \wedge \mathrm{BCFL}_2 = \bigvee_{i=1}^{k-1} \mathcal{G}_{2,2}$, which is included in $\bigvee_{i=1}^{k-1} \mathrm{BCFL}_4 = \mathrm{BCFL}_{4(k-1)}$. This fact implies the containment $\mathcal{G}_{2k-2,2} \vee \mathcal{G}_{2,2} \subseteq \mathrm{BCFL}_{4k}$. It thus follows that $\mathcal{F}_{2k,2m+1} \subseteq \mathrm{BCFL}_{2\ell_1} \vee \mathrm{BCFL}_{2\ell_2} \vee \mathrm{BCFL}_{4k} = \bigvee_{i=1}^{\ell_1+\ell_2+2k} \mathrm{BCFL}_2$. As discussed before, this is equivalent to $\mathrm{BCFL}_{2(\ell_1+\ell_2+2k)}$, which is obviously included in $\mathrm{BCFL}_{\mathrm{BH}}$. Therefore, we conclude that $\mathcal{F}_{2k,2m+1} \subseteq \mathrm{BCFL}_{\mathrm{BH}}$. □

6 Future Challenges

We have initiated a fundamental study on the dissecting power of regular languages and an application of the REG-dissectibility to the i-separation. Throughout our initial study, a number of open questions have arisen for future research. An important open question concerns the REG-dissectibility of co-CFL and, more widely, CFL_k and $\mathrm{CFL}(k)$, which are respectively CFL-analogues of BCFL_k and $\mathrm{BCFL}(k)$, for every level $k \geq 2$. Slightly apart from CFL, two other language families 1-$\mathrm{C}_=$LIN and 1-PLIN, introduced in [8], are, at this moment, unknown to be REG-dissectible. Much anticipated is a development of a coherent theory of a more general notion of $\mathcal{C}$-dissectibility. Concerning the i-separation of $\mathrm{i}(\mathrm{CFL}, \mathrm{CFL})$, on the contrary, a key question of whether CFL i-separate $\mathrm{i}(\mathrm{CFL}, \mathrm{CFL})$ still awaits its answer.

References

[1] Y. Bar-Hillel, M. Perles, and E. Shamir. On formal properties of simple phrase-structure grammars. *Z. Phonetik Sprachwiss. Kommunik.*, 14, 143–172, 1961.

[2] M. Domaratzki, J. Shallit, and S. Yu. Minimal covers of formal languages. In *Proc. of the 5th International Conference on Developments in Language Theory* (DLT 2001), Lecture Notes in Computer Science, Springer, Vol.2295, pp.319–329, 2002.

[3] S. Ginsburg. *The Mathematical Theory of Context-Free Languages.* McGraw-Hill, New York, 1966.

[4] S. Ginsburg and E. H. Spanier. Bounded ALGOL-like languages. *Trans. Amer. Math. Soc.*, 113, 333–368, 1964.

[5] S. Ginsburg and E. H. Spanier. Semigroups, Presburger formulas and languages. *Pacific J. Math.*, 16, 285–296, 1966.

[6] L. Y. Liu and P. Weiner. An infinite hierarchy of intersections of context-free languages. *Math. Systems Theory*, 7, 185–192, 1973.

[7] R. J. Parikh. On context-free languages. *J. ACM*, 13, 570–581, 1961.

[8] K. Tadaki, T. Yamakami, and J. C. H. Lin. Theory of one-tape linear-time Turing machines. *Theor. Comput. Sci.*, 411, 22–43, 2010. An extended abstract appeared in the Proc. of the 30th SOFSEM Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2004), Lecture Notes in Computer Science, Springer, Vol.2932, pp.335–348, 2004.

[9] T. Yamakami. Immunity and pseudorandomness of context-free languages. *Theor. Comput. Sci.*, 412, 6432–6450, 2011.