Kernels of Sub-classes of Context-Free Languages

Martin Kutrib(✉)

Institut für Informatik, Universität Giessen, Arndtstr. 2, 35392 Giessen, Germany
kutrib@informatik.uni-giessen.de

Abstract. While the closure of a language family $\mathscr{L}$ under certain language operations is the least family of languages which contains all members of $\mathscr{L}$ and is closed under all of the operations, a kernel of $\mathscr{L}$ is a greatest family of languages which is a subfamily of $\mathscr{L}$ and is closed under all of the operations. Here we investigate properties of kernels of general language families and operations defined thereon as well as kernels of (deterministic) (linear) context-free languages with a focus on Boolean operations. While the closures of language families usually are unique, this uniqueness is not obvious for kernels. We consider properties of language families and operations that yield unique and non-unique, that is a set, of kernels. For the latter case, the question whether the union of all kernels coincides with the language family, or whether there are languages that do not belong to any kernel is addressed. Furthermore, the intersection of all kernels with respect to certain operations is studied in order to identify sets of languages that belong to all of these kernels.

1 Introduction

Classical and well-developed concepts to represent (formal) languages are, for example, grammars, language equations, or accepting automata. Similarly, families of languages can be represented in several ways. For example, a language family can be defined to be the family of all languages represented by a certain type of grammar, automaton model, language equation, or by applying appropriate operations on other language families. From a practical point of view, there is often a considerable interest in language families that are robust with respect to language operations, that is, the families are preferably closed under the operations, and/or in language families that admit efficient recognizers. A good example are context-free languages, that are one of the most important and most developed area of formal language theory. However, the family is not closed under the two Boolean operations complementation and intersection. Moreover, the known upper bound on the time complexity for context-free language recognition still exceeds $O(n^2)$. As an approach to characterize language families having strong closure properties and efficient recognizers but decrease the expressive capacity only slightly, closures of sub-classes of the context-free languages have been investigated.

A. Chatzigeorgiou et al. (Eds.): SOFSEM 2020, LNCS 12011, pp. 136–147, 2020.
https://doi.org/10.1007/978-3-030-38919-2_12

The Boolean closure of the linear context-free languages offers a significant increase in expressive capacity compared with the linear context-free languages itself. In addition, it preserves the attractively efficient recognition algorithm taking $O(n^2)$ time and $O(n)$ space [11]. In [12], a characterization of deterministic real-time one-way cellular automata by so-called linear conjunctive grammars has been shown. Linear conjunctive grammars are basically linear context-free grammars augmented with an explicit intersection operation, where the number of intersections is, in some sense, not bounded as in a Boolean formula. The systematic investigation of the Boolean closures of arbitrary and deterministic context-free languages started in [14–16], in particular, motivated by the question "How much more powerful is nondeterminism than determinism?" The closure of deterministic languages under the regular operations is studied in [1], while the regular closure of the linear context-free languages is considered in [10].

Here we are interested in language families with strong closure properties obtained by looking into a given family instead of closing and, thus, extending the family. To this end, we study the notion of kernels of language families. Basically, a kernel of some family $\mathscr{L}$ with respect to some language operations defined on $\mathscr{L}$ is a greatest sub-family of $\mathscr{L}$ that is closed under the operations. For example, the family of linear context-free languages is not closed under complementation. Its complementation kernel consists of all linear context-free languages whose complement is also linear context free. This kernel is also known as the family of strongly linear context-free languages that is considered in [8] with respect to its expressive capacity and closure properties. Another question that motivates the concept is as follows. Given a language such that also its complement belongs to the same family, the description of which of both is more economic [8]? For example, it is known that a nondeterministic finite automaton can require 2^n states to accept the complement of a language accepted by an n-state nondeterministic finite automaton [9]. So, a representation of the complement by the n-state automaton together with a bit that says that actually the complement of the language accepted is meant is much more economic from the descriptional complexity point of view. A machine characterization of the complementation kernel of the context-free languages in terms of self-verifying pushdown automata is obtained in [2].

Another well-understood kernel is the family of recursive languages. It is the complementation kernel of the recursively enumerable languages.

The paper is organized as follows. After presenting the basic definitions and notions in the next section, Sect. 3 deals with the uniqueness of kernels. The underlying results are as general as possible while clarifying examples often deal with sub-classes of context-free languages. The question whether any language of a family belongs to some kernel based on given operations is dealt with in Sect. 4. More precisely, we are interested in the question whether the union of all kernels coincides with the language family. The intersection of all of these kernels and its related questions are considered in Sect. 5. Finally, we discuss some interesting untouched problems and questions for further research in Sect. 6.

2 Preliminaries

We write Σ^* for the set of all words over a finite alphabet Σ. The *empty word* is denoted by λ, and we set $\Sigma^+ = \Sigma^* \setminus \{\lambda\}$. The *reversal* of a word w is denoted by w^R, and for the *length* of w we write $|w|$. Set *inclusion* is denoted by $\subseteq$ and *strict set inclusion* by $\subset$.

A subset of Σ^* is called a *(formal) language* over Σ. A *language operation* is an operation whose finite number of parameters are languages, and whose result is a language. For example, the *complement* of a language is defined with respect to the underlying alphabet Σ. For a language $L \subseteq \Sigma^*$, the *complement* $\overline{L}$ of L is $\{\, w \in \Sigma^* \mid w \notin L \,\}$. For all $k \geq 1$, a $k$ary language operation $\circ$ is said to be *idempotent* if $\circ(L, L, \dots, L) = L$, for all L in the domain of $\circ$. For easier writing, here we call even a unary language operation $\circ$ with the property $\circ(L) = L$ idempotent (so we do *not* require $\circ(\circ(L)) = \circ(L)$).

Let Ω be an infinite enumerable set of letters. The set $\mathscr{L}$ is a *family of languages* over Ω if for each $L \in \mathscr{L}$ there is a finite subset $\Sigma \subset \Omega$ such that $L \subseteq \Sigma^*$. In the sequel we tacitly omit Ω when it is understood. For a family of languages $\mathscr{L}$, the family of complements CO-$\mathscr{L}$ is defined to be $\{\, \overline{L} \mid L \in \mathscr{L} \,\}$.

Let $\mathscr{L}$ be a family of languages and $op_1, op_2, \dots, op_k$, $k \geq 1$, be a finite number of operations defined on $\mathscr{L}$.

1. Then $\Gamma_{op_1,op_2,\dots,op_k}(\mathscr{L})$ denotes the $(op_1, op_2, \dots, op_k)$ *closure of* $\mathscr{L}$. That is, the *least family of languages which contains all members of* $\mathscr{L}$ *and is closed under* $op_1, op_2, \dots, op_k$. In other words, there exists no language family $\mathscr{L}'$ that is closed under $op_1, op_2, \dots, op_k$ such that $\mathscr{L} \subseteq \mathscr{L}' \subset \Gamma_{op_1,op_2,\dots,op_k}(\mathscr{L})$.
2. By $\gamma_{op_1,op_2,\dots,op_k}(\mathscr{L})$ we denote the set of $(op_1, op_2, \dots, op_k)$ *kernels of* $\mathscr{L}$. That is, the set of *greatest families of languages which are subfamilies of* $\mathscr{L}$ *and are closed under* $op_1, op_2, \dots, op_k$. In other words, for all kernels $\kappa \in \gamma_{op_1,op_2,\dots,op_k}(\mathscr{L})$ there exists no language family $\mathscr{L}'$ that is closed under $op_1, op_2, \dots, op_k$ such that $\kappa \subset \mathscr{L}' \subseteq \mathscr{L}$.

In particular, we consider the operations complementation ($\sim$), union ($\cup$), and intersection ($\cap$), which are called *Boolean operations.* Accordingly, we write Γ_{BOOL} for $\Gamma_{\sim,\cup,\cap}$ and γ_{BOOL} for$\gamma_{\sim,\cup,\cap}$.

Since special attention is paid to sub-classes of context-free languages, we recall briefly the notion of a context-free grammar and refer to the literature, for example to [7], for detailed definitions of the characterizing automata models.

A *context-free grammar* is a system $G = \langle N, T, S, P \rangle$, where N and T are the disjoint alphabets of nonterminals and terminals, $S \in N$ is the axiom, and P is the finite set of productions of the form $A \to u$, where $A \in N$ and $u \in (N \cup T)^*$. A context-free grammar is said to be *linear* if and only if for all productions the right-hand side u contains at most one nonterminal, that is, $u \in (T^*NT^*) \cup T^*$. A linear grammar is said to be *left-linear* if and only if a nonterminal may only appear as leftmost symbol at the right-hand side of the productions, that is, $u \in (NT^*) \cup T^*$.

The language *generated* by G is the set $\{\, w \in T^* \mid S \Rightarrow^* w \,\}$, where $\Rightarrow^*$ denotes the reflexive, transitive closure of the derivation relation $\Rightarrow$.

The families of languages that can be generated by context-free, linear, and left-linear grammars are called context-free (CFL), linear (LIN), and regular (REG) languages. The automaton model for the recognition of context-free languages is the nondeterministic pushdown automaton. Its deterministic variant characterizes the deterministic context-free languages (DCFL). As for DCFL there is an automaton model for linear languages. Restricting a pushdown automaton such that it may switch from increasing the height of its pushdown to decreasing it only once, thus performing only one turn, leads to the definition of one-turn pushdown automata [5]. It is known that nondeterministic one-turn pushdown automata characterize the linear languages and deterministic one-turn pushdown automata define the deterministic linear languages (DLIN).

3 Uniqueness of Kernels

While the closures of language families under all of the usually considered operations are unique language families, this uniqueness is not obvious for kernels. In fact, it does not always hold. On the other hand, if the kernels are based on unary operations then they are unique, that is, the corresponding set of kernels γ is a singleton.

Proposition 1. *Let $\mathscr{L}$ be a family of languages and $\circ$ be a unary operation defined on $\mathscr{L}$. Then the set $\gamma_\circ(\mathscr{L})$ is a singleton.*

Proof. For any language L from $\mathscr{L}$, the application of $\circ$, that is $\circ(L)$, either does belong to $\mathscr{L}$ or not. Now we consider the iterated application of $\circ$ to $L \in \mathscr{L}$ and define $\circ^1 = \circ$ and, for $1 \leq i$,

$$\circ^{i+1}(L) = \begin{cases} \circ(\circ^i(L)) & \text{if} \quad \circ^i\,(L) \in \mathscr{L} \\ \text{undefined} & \text{else} \end{cases}.$$

So, the iterated application of $\circ$ to languages from $\mathscr{L}$ induces a finite or infinite sequence of (not necessarily different) languages.

If this sequence is finite for some $L \in \mathscr{L}$ then language L does not belong to any $\circ$ kernel of $\mathscr{L}$, since otherwise the kernel would not be closed under $\circ$.

If this sequence is infinite then language L does belong to all $\circ$ kernels of $\mathscr{L}$. If not, all languages $L, \circ^1(L), \circ^2(L), \ldots$ could be added to the kernel without affecting its closure under $\circ$ or its containment in $\mathscr{L}$, a contradiction to the maximality of the kernel.

We conclude that any language from $\mathscr{L}$ either belongs to all $\circ$ kernels or to none $\circ$ kernel. So, the kernel is uniquely determined. □

In general, the uniqueness is lost for $k$ary operations if $k \geq 2$.

Theorem 2. *Let $\mathscr{L}$ be a family of languages, $k \geq 2$, and $\circ$ be a $k$ary idempotent operation defined on $\mathscr{L}$. Then the set $\gamma_\circ(\mathscr{L})$ includes more than one kernel if and only if $\mathscr{L}$ is not closed under $\circ$.*

Proof. If $\mathscr{L}$ is closed under $\circ$, it is its own $\circ$ kernel and, thus, $\gamma_{\circ}(\mathscr{L})$ is a singleton.

Now assume that $\mathscr{L}$ is not closed under $\circ$ and let $L_1, L_2, \dots, L_k \in \mathscr{L}$ be witnesses for the non-closure. That is, $\circ(L_1, L_2, \dots, L_k) \notin \mathscr{L}$. First, we argue that any of the witness languages, say L_i, belongs to a $\circ$ kernel of $\mathscr{L}$. To this end, it suffices to consider the set $\{L_i\}$ which is a subset of $\mathscr{L}$. Since $\circ$ is idempotent the set $\{L_i\}$ is closed under $\circ$. So, either it is a kernel or it is a subset of some kernel.

Now it remains to be concluded that not all of the languages $L_1, L_2, \dots, L_k$ can belong to the same kernel, since this would violate the closure under $\circ$. So, there are at least two different kernels in $\gamma_{\circ}(\mathscr{L})$. □

So far, we obtained that the $\circ$ kernel of some language family is unique if $\circ$ is a unary operation or if the family is closed under $\circ$, and that there are more than one kernels if $\circ$ is a $k$ary idempotent operation, for $k \geq 2$, and $\mathscr{L}$ is not closed under $\circ$. The following examples reveal that a finite as well as an infinite number of kernels may exist.

Example 3. Let $\mathscr{L}$ be defined as union of CFL with $\{L_{\text{expo}}\}$, where L_{expo} is the non-context-free unary language $\{\, a^{2^n} \mid n \geq 0 \,\}$. Family $\mathscr{L}$ is not closed under the idempotent operation union since, for example, $L_{\text{expo}} \cup \{aaa\}$ is not context free and, thus, does not belong to $\mathscr{L}$. By Theorem 2, $\gamma_{\cup}(\mathscr{L})$ includes more than one kernel. In particular, CFL is included in $\gamma_{\cup}(\mathscr{L})$, since CFL is closed under union. This is the only union kernel of $\mathscr{L}$ that does not include L_{expo}.

On the other hand, there must exist a kernel in $\gamma_{\cup}(\mathscr{L})$ having $\{L_{\text{expo}}\}$ as subset, since $\{L_{\text{expo}}\}$ is closed under union and a subset of $\mathscr{L}$. We show that there is exactly one union kernel of $\mathscr{L}$ that includes L_{expo}.

Let $U = \{\, L \mid L \text{ is finite subset of } L_{\text{expo}} \,\}$ be the set of finite languages whose words belong to L_{expo}, and let $R = \{\, L \in \text{CFL} \mid (L \cup L_{\text{expo}}) \cap a^* \in \text{REG} \,\}$ be the set of context-free languages whose unary words from a^* form a regular language when joint with L_{expo}. We claim that $\kappa = U \cup R \cup \{L_{\text{expo}}\}$ is the sole union kernel of $\mathscr{L}$ that includes L_{expo}.

Clearly, we have the inclusion $\kappa \subset \mathscr{L}$. To show that κ is closed under union, let $u, u' \in U$ and $r, r' \in R$. We obtain $u \cup L_{\text{expo}} = L_{\text{expo}} \in \kappa$, $u \cup u' \in U \subset \kappa$, and $u \cup r \in \text{CFL}$, $(u \cup r \cup L_{\text{expo}}) \cap a^* = (r \cup L_{\text{expo}}) \cap a^*$ and, thus $u \cup r \in R \subset \kappa$. Further, we have $r \cup L_{\text{expo}} \cup L_{\text{expo}} = r \cup L_{\text{expo}}$ and, therefore, $r \cup L_{\text{expo}} \in R \subset \kappa$, and $(r \cup r' \cup L_{\text{expo}}) \cap a^* = \big((r \cup L_{\text{expo}}) \cap a^*\big) \cup \big((r' \cup L_{\text{expo}}) \cap a^*\big) \in \text{REG}$ and, thus $r \cup r' \in R \subset \kappa$. We conclude that κ is closed under union.

Finally, it remains to be shown that none of the languages $\mathscr{L} \setminus \kappa$ can belong to any union kernel of $\mathscr{L}$ that includes L_{expo}. This implies that κ is maximal and therefore, in fact, a kernel, and that it is the unique.

So, let $L \in \mathscr{L} \setminus \kappa$. If L includes at least one word that is not of the form a^*, the union $L \cup L_{\text{expo}}$ is not equal to L_{expo}. Since L does not belong to R, we have that $(L \cup L_{\text{expo}}) \cap a^*$ is unary but not regular. So, it is not context free either. Since context-free languages are closed under intersection with regular languages, $L \cup L_{\text{expo}}$ is not context free. It follows that no union kernel of $\mathscr{L}$ that includes L_{expo} includes L.

Next, assume that all words in L are of the form a^*. Since L does not belong to R we now from the previous case that $(L \cup L_{\text{expo}}) \cap a^* = L \cup L_{\text{expo}}$ is not context free. So, if L belongs to the kernel, $L \cup L_{\text{expo}}$ has to be equal to L_{expo}. This implies $L \subseteq L_{\text{expo}}$. Since $L \notin \kappa$ the inclusion is proper: $L \subset L_{\text{expo}}$. Since any infinite subset of L_{expo} is not context free and any finite subset does belong to $U \in \kappa$, we obtain the contradiction that L cannot belong to $\mathscr{L} \setminus \kappa$.

So, we have shown that the set $\gamma_{\cup}(\mathscr{L})$ consists of exactly two kernels, one includes L_{expo} and the other does not. ■

Example 4. The family DLIN is not closed under intersection. We consider the number of kernels in $\gamma_{\cap}(\text{DLIN})$. To this end, for $k \geq 2$, define language $L_k = \{\, a^n(\$a^*)^{k-2}\$a^n(\$a^*)^* \mid n \geq 0 \,\}$ that belongs to DLIN. However, for all $2 \leq i < j$, the intersection $L_i \cap L_j$ is language

$$\{\, a^n(\$a^*)^{i-2}\$a^n(\$a^*)^{j-i-1}\$a^n(\$a^*)^* \mid n \geq 0 \,\}$$

which is not even context free. We conclude that for $2 \leq i, j$ the languages L_i and L_j do not belong to the same kernel if $i \neq j$. On the other hand, for $2 \leq k$, there must exist a kernel in $\gamma_{\cap}(\text{DLIN})$ having $\{L_k\}$ as subset, since it is closed under intersection and a subset of DLIN. So, the set $\gamma_{\cap}(\text{DLIN})$ includes infinitely many kernels. ■

4 Union of Kernels

Next we turn to the question whether any language of a family belongs to some kernel based on given operations. Or are there languages that do not belong to any of such kernels. More precisely, we are interested in the question whether the union of all kernels coincides with the language family.

Theorem 5. *Let $\mathscr{L}$ be a family of languages and $op_1, op_2, \ldots, op_k$, $k \geq 1$, be a finite number of idempotent operations defined on $\mathscr{L}$. Then*

$$\{\, L \mid L \in \kappa \ \textit{ for some } \ \kappa \in \gamma_{op_1, op_2, \ldots, op_k}(\mathscr{L}) \,\} = \mathscr{L}.$$

Proof. The inclusion in $\mathscr{L}$ is trivial. So, it remains to be shown that any language from $\mathscr{L}$ does belong to some $(op_1, op_2, \ldots, op_k)$ kernel of $\mathscr{L}$.

To this end, let $L \in \mathscr{L}$ be an arbitrary language from the family. We consider the set $\nu = \{L\}$. Since it contains only one language and all operations $op_1, op_2, \ldots, op_k$ are idempotent, it is closed under $op_1, op_2, \ldots, op_k$. So, either ν is itself a $(op_1, op_2, \ldots, op_k)$ kernel of $\mathscr{L}$, or there exist a kernel in $\gamma_{op_1, op_2, \ldots, op_k}(\mathscr{L})$ having ν as subset. □

Example 6. Consider the families DLIN, LIN, DCFL, as well as CFL and the idempotent operations union and intersection. Theorem 5 says that any language from one of the families belongs to some $(\cup, \cap)$ kernel of that family. That is,

$$\{\, L \mid L \in \kappa \ \text{ for some } \ \kappa \in \gamma_{\cup,\cap}(\mathscr{L}) \,\} = \mathscr{L},$$

for $\mathscr{L} \in \{\text{DLIN}, \text{LIN}, \text{DCFL}, \text{CFL}\}$. ■

Theorem 5 reveals in particular that idempotent operations do not prevent languages from belonging to a kernel. Let us discuss the role played by the requirement that the operations have to be idempotent. If a unary operation is idempotent, *any* language family is closed under this operation (in fact, the operation is the identity). However, if at least one unary operation under which the family is not closed is in the list, the situation changes.

Proposition 7. *Let $\mathscr{L}$ be a family of languages not closed under the unary operation $\circ$, and $op_1, op_2, \ldots, op_k$, $k \geq 0$, be a finite number of further operations defined on $\mathscr{L}$. Then $\{\, L \mid L \in \kappa$ for some $\kappa \in \gamma_{\circ, op_1, op_2, \ldots, op_k}(\mathscr{L}) \,\} \subset \mathscr{L}$.*

Proof. The inclusion claimed is trivial. So, it remains to be shown that the inclusion is strict.

Since $\mathscr{L}$ is not closed under $\circ$, there is a language $L \in \mathscr{L}$ such that $\circ(L) \notin \mathscr{L}$. So, L cannot belong to any $(\circ, op_1, op_2, \ldots, op_k)$ kernel of $\mathscr{L}$, since the containment would violate the closure of the kernel under $\circ$. □

Example 8. It is well-known that the family CFL is not closed under complementation. Applying Proposition 7 shows that not all context-free languages belong to some Boolean kernel. That is, $\{\, L \mid L \in \kappa$ for some $\kappa \in \gamma_{\mathrm{BOOL}}(\mathrm{CFL}) \,\} \subset$ CFL. ■

In general, the condition of Proposition 7, namely that the family $\mathscr{L}$ has not to be closed under the unary operation, cannot be relaxed. The following proposition shows this fact. It is in contrast to Example 8.

Proposition 9. *Any deterministic context-free language belongs to some kernel $\kappa \in \gamma_{BOOL}(\mathrm{DCFL})$.*

Proof. Let $L \in$ DCFL be some language over the alphabet Σ. We consider the set $\nu = \{L, \overline{L}, \Sigma^*, \emptyset\}$ which is clearly closed under complementation, union, and intersection.

Since DCFL is closed under complementation and includes the regular languages Σ^* and $\emptyset$, either ν is itself a Boolean kernel of DCFL, or there exists a kernel in $\gamma_{\mathrm{BOOL}}(\mathrm{DCFL})$ having ν, and thus $\{L\}$, as subset. □

In order to continue the discussion of the requirement that the operations have to be idempotent, we present a further example considering the binary non-idempotent operation of marked concatenation.

Example 10. The family LIN is not closed under the binary non-idempotent operation of marked concatenation ($\bullet$). In fact, it has been shown in [6] that the marked concatenation of two linear context-free languages is linear context free if and only if at least one of the languages is regular.

We consider $\gamma_\bullet(\mathrm{LIN})$. Since the family REG is closed under marked concatenation, there must be some $\kappa \in \gamma_\bullet(\mathrm{LIN})$ such that REG $\subseteq \kappa$. On the other hand, let $L \in \mathrm{LIN} \setminus \mathrm{REG}$ be an arbitrary linear context-free language that is not regular. Then L cannot belong to any kernel in $\gamma_\bullet(\mathrm{LIN})$ since $L \bullet L$ is not linear context free due to [6]. Therefore, REG is the sole marked concatenation kernel of LIN. That is, $\gamma_{\bullet}(\text{LIN}) = \{\text{REG}\}$ and, thus, the marked concatenation kernel of LIN is unique. Moreover, $\{\, L \mid L \in \kappa \text{ for some } \kappa \in \gamma_{\bullet}(\text{LIN}) \,\} \subset \text{LIN}$. ■

It is worth mentioning that literally Example 10 also applies to the family DLIN.

5 Intersection of Kernels

We now turn to the question which languages belong to all kernels based on given operations. So, we consider the intersection of all of these kernels.

Proposition 11. *Let $\mathscr{L} \in \{\text{CFL}, \text{LIN}, \text{DCFL}, \text{DLIN}\}$. All intersection kernels and union kernels of $\mathscr{L}$ include* REG.

Proof. In contrast to the assertion assume that there is a kernel $\nu \in \gamma_{\cap}(\mathscr{L})$ such that $\text{REG} \not\subseteq \nu$.

In order to obtain a contradiction we show that ν is strictly included in a kernel from $\gamma_{\cap}(\mathscr{L})$ and, thus, cannot be an intersection kernel of $\mathscr{L}$ at all. To this end, we join ν with REG and build the intersection closure of the union. That is, we consider $\kappa = \Gamma_{\cap}(\nu \cup \text{REG})$.

Any language $L \in \kappa$ has a representation of the form K, R, or $K \cap R$, where $K \in \nu$ and $R \in \text{REG}$. Since $\mathscr{L}$ includes the regular languages and is closed under intersection with regular languages, language L belongs to $\mathscr{L}$. So, we have $\Gamma_{\cap}(\nu \cup \text{REG}) \subseteq \mathscr{L}$. This shows the assertion for intersection kernels.

Since $\mathscr{L}$ is closed under union with regular languages as well, the argumentation for union kernels follows by replacing intersection with union. □

Of particular interest are the languages that belong to *all* Boolean kernels.

Theorem 12. *Let $\mathscr{L} \supseteq \mathscr{T}$ be two families of languages. If $\mathscr{L}$ is closed under union and under intersection with languages from $\mathscr{T}$, and $\mathscr{T}$ is closed under the Boolean operations then $\mathscr{T} \subseteq \kappa$ for all $\kappa \in \gamma_{BOOL}(\mathscr{L})$.*

Proof. In contrast to the assertion assume that there is a kernel $\nu \in \gamma_{\text{BOOL}}(\mathscr{L})$ such that $\mathscr{T} \not\subseteq \nu$.

In order to obtain a contradiction we show that ν is strictly included in a kernel from $\gamma_{\text{BOOL}}(\mathscr{L})$ and, thus, cannot be a Boolean kernel of $\mathscr{L}$ at all. To this end, we join ν with $\mathscr{T}$ and build the Boolean closure of the union. That is, we consider $\kappa = \Gamma_{\text{BOOL}}(\nu \cup \mathscr{T})$. We show that κ is included in $\mathscr{L}$.

Let $L \in \kappa$. Then, for some $m, l_1, l_2, \dots, l_m \geq 0$, language L has a representation $\bigcup_{1 \leq i \leq m} \bigcap_{1 \leq j \leq l_i} L_{i,j}$ such that $L_{i,j} \in (\nu \cup \mathscr{T})$ or $L_{i,j} \in \text{CO-}(\nu \cup \mathscr{T})$. Since ν as well as $\mathscr{T}$ are closed under complementation, we have $(\nu \cup \mathscr{T}) = \text{CO-}(\nu \cup \mathscr{T})$, and may safely assume that $L_{i,j} \in (\nu \cup \mathscr{T})$.

Now, for $1 \leq i \leq m$, let $L_i = L_{i,1} \cap L_{i,2} \cap \cdots \cap L_{i,l_i}$. Since ν as well as $\mathscr{T}$ are closed under intersection, we have $L_i = K_i \cap T_i$ or $L_i = K_i$ or $L_i = T_i$, for some

$K_i \in \nu$ and $T_i \in \mathscr{T}$. Moreover, since ν and $\mathscr{T}$ are sub-families of $\mathscr{L}$, and $\mathscr{L}$ is closed under intersection with languages from $\mathscr{T}$, language L_i belongs to $\mathscr{L}$.

Finally, $L = \bigcup_{1 \leq i \leq m} L_i$ and the closure of $\mathscr{L}$ under union implies that L belongs to $\mathscr{L}$. Therefore, κ is included in $\mathscr{L}$. □

Corollary 13. *Let $\mathscr{L} \supseteq \mathscr{T}$ be two families of languages. If $\mathscr{L}$ is closed under intersection and under union with languages from $\mathscr{T}$, and $\mathscr{T}$ is closed under the Boolean operations then $\mathscr{T} \subseteq \kappa$ for all $\kappa \in \gamma_{BOOL}(\mathscr{L})$.*

Proof. The corollary can be shown almost literally as Theorem 12, where the representation of language $L \in \kappa$ is given as $\bigcap_{1 \leq i \leq m} \bigcup_{1 \leq j \leq l_i} L_{i,j}$, and by interchanging union and intersection in the reasoning. □

Example 14. The families CFL and LIN are closed under union and under intersection with regular languages. The family of regular languages is closed under the Boolean operations. So, by applying Theorem 12 we obtain that *all* Boolean kernels of CFL and LIN include REG.

Moreover, applying Corollary 13 shows that *all* Boolean kernels of CO-CFL and CO-LIN include REG. ■

Since any intersection, union, and complementation kernel of CFL, LIN, CO-CFL, and CO-LIN includes a Boolean kernel which, in turn, includes REG, all of these kernels include REG as well. Moreover, for all unary operations $\circ$ under which the family of regular languages is closed, the unique $\circ$ kernel of CFL, LIN, CO-CFL, and CO-LIN includes REG (see Proposition 1). This immediately raises the question whether these kernels are characterized by REG. Or are there certain non-regular languages that belong to *all* kernels of a certain type. Example 10 shows that REG is the sole marked concatenation kernel of LIN and, thus, characterizes the kernel. However, in the following we turn to show that there are non-regular languages belonging to the intersection of all Boolean kernels of CFL, LIN, CO-CFL, and CO-LIN.

To this end, we recall the notion of semilinear languages. Consider, for some fixed positive integer m, the vectors in $\mathbb{N}^m$. A set of the form

$$\{\, v_0 + x_1 v_1 + x_2 v_2 + \cdots + x_k v_k \mid x_i \geq 0, 1 \leq i \leq k \,\},$$

where $v_0, v_1, \ldots, v_k \in \mathbb{N}^m$, is said to be *linear*. A *semilinear* set is a finite union of linear sets. It is known that the family of semilinear subsets of $\mathbb{N}^m$ is closed under union, intersection, and complementation [3]. For an alphabet $\Sigma = \{a_1, a_2, \ldots, a_m\}$ the *Parikh mapping* $\Psi \colon \Sigma^* \to \mathbb{N}^m$ is defined by $\Psi(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_m})$, where $|w|_{a_i}$ denotes the number of occurrences of a_i in the word w. In [13] a fundamental result concerning the distribution of symbols in the words of a context-free language has been shown. It says that for any context-free language L, the Parikh image $\Psi(L) = \{\, \Psi(w) \mid w \in L \,\}$ is semilinear.

In the following we consider semilinear languages that are subsets of a^*b^*, where the number of b's depends linearly on the number of a's. The dependency is given by linear functions $\varphi\colon \mathbb{N} \to \mathbb{N}$ with $\varphi(n) = c_1 \cdot n + c_0$, for some $c_0, c_1 \geq 0$. So, we define $L_\varphi = \{\, a^n b^{\varphi(n)} \mid n \geq 0 \,\}$. Note that there are functions φ such that L_φ is context free but not regular (for example $\varphi(n) = n$, $\varphi(n) = 2n$, etc.), or L_φ is regular (for example $\varphi(n)$ is constant). However, the linearity of φ implies that L_φ is a semilinear language, where $\Psi(L_\varphi) = \left\{ \binom{0}{c_0} + x\binom{1}{c_1} \,\middle|\, x \geq 0 \right\}$.

Theorem 15. *Let $\varphi\colon \mathbb{N} \to \mathbb{N}$ be a linear function. For an arbitrary context-free language L, the intersection $L \cap L_\varphi$ belongs to* DLIN*.*

Proof. We consider the Parikh image

$$S = \Psi(L \cap L_\varphi) = \Psi((L \cap a^*b^*) \cap L_\varphi) = \Psi(L \cap a^*b^*) \cap \Psi(L_\varphi).$$

The set S is semilinear since $L \cap a^*b^*$ is context free and, thus, semilinear [13], language L_φ is semilinear, and semilinear sets are closed under intersection [3].

Let $\pi_1\colon \mathbb{N}^2 \to \mathbb{N}$ be the canonical projection on the first factor. Then $\pi_1(S)$ is semilinear. So, the language $U = \{\, a^n \mid n \geq 0, a^n b^{\varphi(n)} \in L \,\} = \Psi^{-1}(\pi_1(S))$ is regular since it is unary and semilinear.

Now, let M be a deterministic finite automaton accepting U. From M one can easily construct a deterministic one-turn pushdown automaton accepting $\{\, a^n b^{\varphi(n)} \mid n \geq 0, a^n \in U \,\} = L \cap L_\varphi$. So, the theorem follows. □

Example 16. Let $\varphi\colon \mathbb{N} \to \mathbb{N}$ be a linear function. Then, for all families $\mathscr{L}$ from $\{\mathrm{CFL}, \mathrm{LIN}, \mathrm{DCFL}, \mathrm{DLIN}\}$, all intersection kernels of $\mathscr{L}$ include all, even non-regular, languages L_φ.

Similar as above we obtain a contradiction when we assume that there is an intersection kernel $\nu \in \gamma_\cap(\mathscr{L})$ such that there is $L_\varphi \notin \nu$.

Consider $\kappa = \Gamma_\cap(\nu \cup \{L_\varphi\})$. Each language $L \in \kappa$ has a representation as K, L_φ, or $K \cap L_\varphi$, where $K \in \nu$.

Since L_φ belongs to $\mathscr{L}$, $K \cap L_\varphi \in \mathrm{DLIN} \subseteq \mathscr{L}$ by Theorem 15, and $\nu \subseteq \mathscr{L}$, the closure $\Gamma_\cap(\nu \cup \{L_\varphi\})$ is included in $\mathscr{L}$, which gives a contradiction to the maximality of ν. ■

The situation changes when in Theorem 15 the language L_φ is replaced by its complement $\overline{L_\varphi}$. It is an immediate observation that in this case the determinism is not generally achieved. However, we can show that the property of being context free or linear context free can be preserved. To this end, we first provide the next lemma.

It has already been shown in [4] that a language $L \subseteq a^*b^*$ is context free if and only if it is semilinear. We turn to strengthen this result to linear context-free languages. Basically, it shows that there are no non-linear context-free languages $L \subseteq a^*b^*$ at all.

Proposition 17. *A language $L \subseteq a^*b^*$ is linear context free if and only if it is semilinear.*

Proof. If language L is linear context free, it is semilinear. So, it is sufficient to show the converse. To this end, let $L \subseteq a^*b^*$ be semilinear. A semilinear subset S of $\mathbb{N}^2$ determines uniquely a language $\Psi^{-1}(S)$ whose words are of the form a^*b^*, that is $L = \Psi^{-1}(\Psi(L))$. Now let the Parikh image $\Psi(L)$ be given by a finite union of sets of the form

$$\left\{ \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + x_1 \begin{pmatrix} u_1 \\ v_1 \end{pmatrix} + x_2 \begin{pmatrix} u_2 \\ v_2 \end{pmatrix} + \cdots + x_k \begin{pmatrix} u_k \\ v_k \end{pmatrix} \,\middle|\, x_i \geq 0, 1 \leq i \leq k \right\},$$

where $u_0, v_0, u_1, v_1, \ldots, u_k, v_k \in \mathbb{N}$.

For each of these sets, say set S', we construct a linear context-free grammar that generates $\Psi^{-1}(S')$. Since the family of linear context-free languages is closed under union, this shows the lemma.

The linear context-free grammar for S' is $G = \langle N, T, A, P \rangle$, where $N = \{A\}$, $T = \{a, b\}$, and $P = \{\, A \to a^{u_i} A b^{v_i} \mid 1 \leq i \leq k \,\} \cup \{A \to a^{u_0} b^{v_0}\}$. □

Theorem 18. *Let* $\varphi\colon \mathbb{N} \to \mathbb{N}$ *be a linear function,* $\mathscr{L} \in \{\mathrm{CFL}, \mathrm{LIN}\}$*, and* $L \in \mathscr{L}$ *be arbitrary. Then the intersection* $L \cap \overline{L_\varphi}$ *belongs to* $\mathscr{L}$*.*

Proof. The intersection $L \cap \overline{L_\varphi}$ consists of all words from L that are not of the form a^*b^*, and all words from L of the form a^*b^* where the number of b's is different from φ applied to the number of a's. So, we have the representation $L \cap \overline{L_\varphi} = (L \setminus a^*b^*) \cup ((L \cap a^*b^*) \setminus L_\varphi)$.

Since $\mathscr{L}$ is closed under set difference with regular languages, $L \setminus a^*b^*$ belongs to $\mathscr{L}$. Since $\mathscr{L}$ is closed under intersection with regular languages, $L \cap a^*b^*$ belongs to $\mathscr{L}$ and, thus, is semilinear. Further, L_φ is semilinear. The family of semilinear languages is closed under set difference [3]. Therefore, $(L \cap a^*b^*) \setminus L_\varphi$ is a semilinear language which, in turn, is linear context free by Proposition 17 and, thus, belongs to $\mathscr{L}$ as well.

Since $\mathscr{L}$ is closed under union, the intersection $L \cap \overline{L_\varphi}$ belongs to $\mathscr{L}$. □

Now we are prepared to show that there are non-regular languages belonging to the intersection of all Boolean kernels of CFL, LIN, CO-CFL, and CO-LIN.

Theorem 19. *Let* $\varphi\colon \mathbb{N} \to \mathbb{N}$ *be a linear function. Then, for all families* $\mathscr{L}$ *from* $\{\mathrm{CFL}, \mathrm{LIN}, \text{CO-CFL}, \text{CO-LIN}\}$*, all Boolean kernels of* $\mathscr{L}$ *include all, even non-regular, languages* L_φ*.*

6 Untouched Questions

We have started to study the properties of kernels of general language families and operations defined thereon systematically as well as kernels of (deterministic) (linear) context-free languages with a focus on Boolean operations.

Since only less is known about kernels a bunch of questions and problems remain open or untouched. Exemplarily, we mention four of them: (1) The non-trivial closure properties of kernels themselves are of natural interest. (2) Are there hierarchies of kernels? (3) A machine characterization of the complementation kernel of the context-free languages in terms of self-verifying pushdown automata is known [2]. Basically, the characterization is given by a machine for the underlying language family, where the acceptance condition is modified. Are there machine characterizations of other kernels?

Acknowledgment. The author would like to thank Henning Fernau for fruitful discussions at an early stage of the paper.

References

1. Bertsch, E., Nederhof, M.J.: Regular closure of deterministic languages. SIAM J. Comput. **29**, 81–102 (1999)
2. Fernau, H., Kutrib, M., Wendlandt, M.: Self-verifying pushdown automata. In: Non-Classical Models of Automata and Applications (NCMA 2017), vol. 329, pp. 103–117. Austrian Computer Society, Vienna (2017). books@ocg.at
3. Ginsburg, S.: The Mathematical Theory of Context-Free Languages. McGraw Hill, New York (1966)
4. Ginsburg, S., Spanier, E.H.: Bounded ALGOL-like languages. Trans. Am. Math. Soc. **113**, 333–368 (1964)
5. Ginsburg, S., Spanier, E.H.: Finite-turn pushdown automata. SIAM J. Contr. **4**, 429–453 (1966)
6. Greibach, S.A.: The unsolvability of the recognition of linear context-free languages. J. ACM **13**, 582–587 (1966)
7. Harrison, M.A.: Introduction to Formal Language Theory. Addison-Wesley, Reading (1978)
8. Ilie, L., Păun, G., Rozenberg, G., Salomaa, A.: On strongly context-free languages. Discrete Appl. Math. **103**, 158–165 (2000)
9. Jirásková, G.: State complexity of some operations on binary regular languages. Theoret. Comput. Sci. **330**, 287–298 (2005)
10. Kutrib, M., Malcher, A.: Finite turns and the regular closure of linear context-free languages. Discrete Appl. Math. **155**, 2152–2164 (2007)
11. Kutrib, M., Malcher, A., Wotschke, D.: The Boolean closure of linear context-free languages. Acta Inform. **45**, 177–191 (2008)
12. Okhotin, A.: Automaton representation of linear conjunctive languages. In: Ito, M., Toyama, M. (eds.) DLT 2002. LNCS, vol. 2450, pp. 393–404. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45005-X_35
13. Parikh, R.J.: On context-free languages. J. ACM **13**, 570–581 (1966)
14. Wotschke, D.: Nondeterminism and Boolean operations in PDA's. J. Comput. Syst. Sci. **16**, 456–461 (1978)
15. Wotschke, D.: The Boolean closures of the deterministic and nondeterministic context-free languages. In: Brauer, W. (ed.) GI 1973. LNCS, vol. 1, pp. 113–121. Springer, Heidelberg (1973). https://doi.org/10.1007/3-540-06473-7_11
16. Wotschke, D.: Degree-languages: a new concept of acceptance. J. Comput. Syst. Sci. **14**(2), 187–209 (1977)