Unary Context-Free Languages are Regular

Theo Teske

October 14, 2025

1 Introduction

Normally, the class of regular languages is a strict subset of the class of context-free languages. However, when the alphabet size is $|\Sigma| = 1$, the distinction between regular languages and context-free languages vanishes. This phenomenon is very closely related to Parikh's theorem, which essentially states that if we ignore the order of symbols and only count their frequencies, then every context-free language looks regular.

Definition 1 (Parikh vector)**.** The Parikh vector of a string w is defined as its image under the function $\psi : \Sigma^* \to \mathbb{N}^k$ given by

$$\psi(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_k}),$$

where $|w|_{a_i}$ denotes the number of occurrences of the symbol a_i in the string w.

Theorem 1 (Parikh's theorem)**.** *Let L be a context-free language. Then the set of Parikh vectors of strings in L, that is, $\{\psi(w) \mid w \in L\}$, is semi-linear.*

However, the full force of Parikh's theorem is not necessary to explore this unary collapse of the first two levels of the Chomsky hierarchy, and herein we give an alternative route to arrive at this remarkable conclusion.

2 Proofs

First, we prove what is essentially Parikh's theorem in one dimension by showing that for any context-free language L over a unary alphabet, the set of Parikh vectors of all strings in L is the union of finitely many arithmetic progressions.

Theorem 2. *Let $L \subseteq \{a\}^*$ be a context-free language over a unary alphabet. Then the set of string-lengths $S = \{n \in \mathbb{N} \mid a^n \in L\}$ is semilinear.*

Proof. Since L is a context-free language, the Pumping Lemma for context-free languages applies. There exists a pumping length $p \in \mathbb{N}$ such that for any string $w \in L$ such that $|w| \geq p$, w can be written as $w = uvxyz$ satisfying

1. For each $k \geq 0$, $uv^kxy^kz \in L$,
2. $|vy| > 0$, and
3. $|vxy| \leq p$.

Since the alphabet is unary, the positions of the substrings v and y do not affect the resulting string, only their combined length. Let $n = |w|$ and let $d = |v| + |y|$. The pumping condition implies that if $n \in S$ and $n \geq p$, then there exists an integer d such that $1 \leq d \leq p$ and $n + kd \in S$ for all $k \geq 0$.

Let $S_{\geq p} := \{n \in S \mid n \geq p\}$ be the set of lengths of words in L greater than or equal to the pumping length. For every $n \in S_{\geq p}$, there exists at least one valid pumping increment $d \in \{1, \ldots, p\}$. We collect these valid increments into a relation R defined by

$$R := \{(n, d) \in S_{\geq p} \times \{1, \ldots, p\} \mid \forall k \geq 0, (n + kd) \in S\}.$$

We can partition $S_{\geq p}$ based on these increments. For each possible increment $d \in \{1, \ldots, p\}$, define the set of lengths that allow pumping by d as

$$Q_d := \{n \in S_{\geq p} \mid (n, d) \in R\}.$$

Since every $n \geq p$ must have at least one valid pumping length, we have

$$S_{\geq p} = \bigcup_{d=1}^{p} Q_d.$$

We now decompose each Q_d into arithmetic progressions based on residues modulo d. For a fixed d and a residue $r \in \{0, 1, \ldots, d-1\}$, let $Q_{d,r}$ be the subset of lengths in Q_d congruent to r modulo d, so

$$Q_{d,r} := \{n \in Q_d \mid n \equiv r \pmod{d}\}.$$

If $Q_{d,r}$ is empty, it contributes nothing to the union. If it is non-empty, let $m_{d,r} := \min(Q_{d,r})$ denote the minimum element in this subset.

Construct the arithmetic progression starting at $m_{d,r}$ with step d as

$$A_{d,r} := \{m_{d,r} + kd \mid k \in \mathbb{N}\}.$$

We assert that $S_{\geq p}$ is exactly equal to the union of these progressions.

$\subseteq$: By definition, $m_{d,r} \in Q_d$, which means that $(m_{d,r}, d) \in R$. By the definition of R, the entire sequence generated by pumping $m_{d,r}$ with step d is contained in S. Thus, $A_{d,r} \subseteq S$.

$\supseteq$: Take any $n \in S_{\geq p}$. It must belong to some Q_d, so it must also belong to some $Q_{d,r}$ where $r \equiv n \bmod d$. Since $m_{d,r}$ is the minimum element of that set, $n \geq m_{d,r}$. Finally, because $n \equiv m_{d,r} \bmod d$, n can be written as $m_{d,r} + kd$ for some $k \in \mathbb{N}$. Thus, $n \in A_{d,r}$.

Let $F = \{n \in S \mid n < p\}$ be the finite set of small lengths in S. Then by the above,

$$S = F \cup \bigcup_{d=1}^{p} \bigcup_{r=0}^{d-1} A_{d,r}.$$

Since F is finite and the union over d and r is a finite union of arithmetic progressions, S is semi-linear, as desired. □

Remark. The application of the pumping lemma in the preceding proof exposes the underlying mathematical reason why Theorem 2 is true, and this provides insight into Parikh's theorem as well. In general, string concatenation is not commutative, as for instance, $ab \neq ba$, so ignoring the order of symbols in a string loses information. However, in a unary CFL, meaning a CFL over a unary alphabet, concatenation is indeed commutative; for example, $a^3a^2 = a^2a^3$, a fact upon which we relied. Thus, in this case ignoring order loses nothing, and the Parikh image (meaning the set of string-lengths S) describes the language completely.

By proving that the string-lengths of a unary CFL form arithmetic progressions, we showed that beyond a certain threshold, the language's structure is merely counting modulo some period M. This is formalized below.

Theorem 3. *Let $L \subseteq \{a\}^*$ be a unary context-free language. Then the set of string-lengths $S = \{n \in \mathbb{N} \,|\, a^n \in L\}$ is ultimately periodic: there exist integers $N_0, M \in \mathbb{N}$ such that for all $n \geq N_0$,*

$$n \in S \iff n + M \in S.$$

Proof. By Theorem 2, the set of lengths S can be expressed as

$$S = F \cup \bigcup_{i=0}^{k} \{m_i + kd_i \mid k \geq 0\}$$

where F is a finite set of integers and the union represents a finite number of arithmetic progressions.

Define the period M to be the least common multiple of all common differences d_i, so

$$M := \operatorname{lcm}\{d_i \mid 1 \leq i \leq k\},$$

and choose the threshold N_0 to be large enough to clear the finite set and the start points of all arithmetic progressions, so

$$N_0 := \max(\{\max(F) + 1\} \cup \{m_i \mid 1 \leq i \leq k\}).$$

$\Longrightarrow$: Let $n \geq N_0$ and $n \in S$. Since $n \geq \max(F) + 1$, n must belong to the union of arithmetic progressions. Thus, there is some m_i and d_i such that $n = m_i + kd_i \in S$ for all $k \geq 0$. Since d_i divides M, we can write $M = jd_i$ for some $j \in \mathbb{N}$, and therefore

$$n + M = (m_i + kd_i) + jd_i = m_i + (k + j)d_i \in S.$$

$\Longleftarrow$: Let $n \geq N_0$ and $n + M \in S$. Since $n + M \geq n \geq \max(F) + 1$, $n + M$ must belong to the union of arithmetic progressions, and there is some m_i and d_i such that $n + M = m_i + kd_i \in S$ for all $k \geq 0$. As d_i divides M, we can write $M = jd_i$ for some $j \in \mathbb{N}$, and thus

$$n = n + M - M = (m_i + kd_i) - jd_i = m_i + (k - j)d_i.$$

Now, we need only show that $(k - j) \geq 0$, but this follows immediately from the fact that $n \geq N_0 \geq m_i$. So, $n \in S$.

We conclude that S is ultimately periodic with period M and threshold N_0. □

Remark. Interestingly, Theorem 3 proves that context-free grammars cannot "count" in complex ways (like powers or primes). Rather, they can only count linearly. For example, the language of squares

$$L = \{a^{n^2} \mid n \geq 0\} = \{\varepsilon, a, a^4, a^9, a^{16}, \dots\}$$

is not context-free by Theorem 3 because the gaps between consecutive lengths grow infinitely and never become periodic. Similarly, the language of primes

$$L = \{a^p \mid p \text{ is prime}\}$$

is not context-free because the gaps between consecutive prime numbers are not periodic (there are arbitrarily large gaps between successive primes).

We finally apply the elegant characterization of regular languages provided by the Myhill-Nerode theorem in combination with Theorem 3 to arrive at the desired result.

Definition 2 (Myhill-Nerode relation)**.** Let x and y be strings and let L be any language. We write $x \equiv_L y$ if for every string z,

$$xz \in L \iff yz \in L.$$

Theorem 4 (Myhill-Nerode theorem)**.** *Let L be a language. Then L is regular if and only if its Myhill-Nerode relation $\equiv_L$ partitions the set of all strings Σ^* into a finite number of equivalence classes.*

Theorem 5. *If $L \subseteq \{a\}^*$ is context-free, then L is regular.*

Proof. Let $S = \{n \in \mathbb{N} \,|\, a^n \in L\}$. By the previous theorem, S is ultimately periodic with parameters N_0 and M: for all $n \geq N_0$, we have

$$n \in S \iff n + M \in S.$$

Observe that $a^i \equiv_L a^j$ if and only if for all $k \geq 0$,

$$a^{i+k} \in L \iff a^{j+k} \in L,$$

where $\equiv_L$ is the Myhill-Nerode relation of L.

This means that $a^i \equiv_L a^j$ precisely when the tail sets $\{k \mid i + k \in S\}$ and $\{k \mid j + k \in S\}$ are identical.

For $i \geq N_0$, the tail set depends only on $i \bmod M$ (by ultimate periodicity). Therefore, there are at most N_0 equivalence classes for $i < N_0$, and there are at most M equivalence classes for $i \geq N_0$.

Thus, $\equiv_L$ has at most $N_0 + M$ equivalence classes, and so L is regular by the Myhill-Nerode theorem. □