Regularity of non context-free languages over a singleton terminal alphabet

Alberto Pettorossi[a,b,*], Maurizio Proietti[b]

[a] *DICII, University of Rome Tor Vergata, Rome, Italy*
[b] *CNR-IASI, Rome, Italy*

Abstract

It is well-known that: (i) every context-free language over a singleton terminal alphabet is regular [4], and (ii) the class of languages that satisfy the Pumping Lemma is a proper super-class of the context-free languages. We show that any language in this super-class over a singleton terminal alphabet is regular. Our proof is based on a transformational approach and does not rely on Parikh's Theorem [6]. Our result extends previously known results because there are languages that are not context-free, do satisfy the Pumping Lemma, and do not satisfy the hypotheses of Parikh's Theorem [7].

Keywords: Context-free languages, pumping lemma, Parikh's Theorem, regular languages.

Let us begin by introducing our terminology and notations.

The set of the natural numbers is denoted by N. The set of the n-tuples of natural numbers is denoted by N^n. We say that a language L is over the terminal alphabet Σ iff $L \subseteq \Sigma^*$. Given a word $w \in \Sigma^*$, w^0 is the empty word ε, and, for any $i \geq 0$, w^{i+1} is $w^i\, w$, that is, the concatenation of w^i and w. The length of a word w is denoted by $|w|$. Given a symbol $a \in \Sigma$, the number of occurrences of a in w is denoted by $|w|_a$. The cardinality of a set A is denoted by $|A|$.

Given an alphabet Σ such that $|\Sigma| = 1$, the concatenation of any two words w_1, w_2 in Σ^* is commutative, that is, $w_1\, w_2 = w_2\, w_1$.

*Corresponding author.

Email addresses: `pettorossi@info.uniroma2.it` (Alberto Pettorossi), `maurizio.proietti@iasi.cnr.it` (Maurizio Proietti)

In Theorem 2 below we extend the well known result stating that any context-free language over a singleton terminal alphabet is a regular language [4]. An early proof of this result appears in a paper by Ginsburg and Rice [3]. That proof is based on Tarski's fixpoint theorem and it is not based on the Pumping Lemma (contrary to what has been stated in a paper by Andrei et al. [1]). Our extension is due to the facts that: (i) our proof does not rely on Parikh's Theorem [6], like the proof in Harrison's book [4], and (ii) there are non context-free languages that do satisfy the Pumping Lemma (see Definition 1) and do not satisfy Parikh's Condition (see Definition 2) (and thus Parikh's Theorem cannot be applied) [7].

Definition 1 (Pumping Lemma [2]). We say that a language $L \subseteq \Sigma^*$ satisfies the Pumping Lemma iff the following property, denoted $PL(L)$, holds: $\exists\, n > 0, \forall\, z \in L$, if $|z| \geq n$, then $\exists\, u, v, w, x, y \in \Sigma^*$, such that

(1) $z = u\,v\,w\,x\,y$,
(2) $v\,x \neq \varepsilon$,
(3) $|v\,w\,x| \leq n$, and
(4) $\forall\, i \geq 0$, $u\,v^i\,w\,x^i\,y \in L$. □

Definition 2 (Parikh's Condition [6]). (i) A subset S of N^n is said to be a *linear* set iff there exist $v_0, \ldots, v_k \in N^n$ such that $S = \{v_0 + n_1\,v_1 + \ldots + n_k\,v_k \mid n_1, \ldots, n_k \in N\}$, where, for any given $u = \langle u_1, \ldots, u_n \rangle$ and $v = \langle v_1, \ldots, v_n \rangle$ in N^n, $u + v$ denotes $\langle u_1 + v_1, \ldots, u_n + v_n \rangle$ and, for any $m \geq 0$, $m\,u$ denotes $\langle m\,u_1, \ldots, m\,u_n \rangle$. (ii) Given the alphabet $\Sigma = \{a_1, \ldots, a_n\}$, we say that a language $L \subseteq \Sigma^*$ satisfies Parikh's Condition iff $\{\langle |w|_{a_1}, \ldots, |w|_{a_n} \rangle \mid w \in L\}$ is a finite union of linear subsets of N^n. □

Let us first state and prove the following lemma whose proof is by transformation from Definition 1.

Lemma 1. For any language L over a terminal alphabet Σ such that $|\Sigma| = 1$, $PL(L)$ holds iff the following property, denoted $PL1(L)$, holds:
$\exists\, n > 0, \forall\, z \in L$, if $|z| \geq n$, then $\exists\, p \geq 0$, $q \geq 0$, $m \geq 0$, such that

(1.1) $|z| = p + q$,
(2.1) $0 < q \leq n$,
(3.1) $0 < m + q \leq n$, and
(4.1) $\forall\, s \in \Sigma^*, \forall\, i \geq 0$, if $|s| = p + i\,q$, then $s \in L$.

PROOF. If $|\Sigma|=1$, then commutativity of concatenation implies that in $PL(L)$ we can replace $v\,w\,x$ by $w\,v\,x$, and $u\,v^i\,w\,x^i\,y$ by $u\,y\,w\,(v\,x)^i$. Then, we can replace: $u\,y$ by $\widetilde{u}$, $v\,x$ by $\widetilde{v}$, and $(\exists\, u,\ v,\ w,\ x,\ y)$ by $(\exists\,\widetilde{u},\ \widetilde{v},\ w)$. Thus, from $PL(L)$, we get: $\exists\, n>0$, $\forall\, z \in L$, if $|z|\geq n$, then $\exists\,\widetilde{u},\widetilde{v}, w \in \Sigma^*$, such that

(1′) $z = \widetilde{u}\, w\, \widetilde{v}$,

(2′) $\widetilde{v} \neq \varepsilon$,

(3′) $|w\,\widetilde{v}| \leq n$, and

(4′) $\forall\, i\geq 0$, $\widetilde{u}\, w\, \widetilde{v}^i \in L$.

Now if we take the lengths of the words and we denote $|\widetilde{u}\, w|$ by p, $|\widetilde{v}|$ by q, and $|w|$ by m, we get:

$\exists\, n>0$, $\forall\, z \in L$, if $|z|\geq n$, then $\exists\, p\geq 0$, $q\geq 0$, $m\geq 0$, such that

(1″) $|z| = p + q$,

(2″) $q>0$,

(3″) $m+q \leq n$, and

(4″) $\forall\, s\in\Sigma^*$, $\forall\, i\geq 0$, if $|s| = p + i\,q$, then $s \in L$.

For all $n > 0$, $q > 0$, and $m \geq 0$, we have that $(q > 0 \wedge m+q \leq n)$ iff $(0<q\leq n \ \wedge\ 0<m+q\leq n)$. Thus, we get $PL1(L)$. $\square$

We say that $PL1(L)$ *holds for* b if b is a witness of the quantification '$\exists\, n>0$' in $PL1(L)$. The following theorem states our main result.

Theorem 2. *Let L be any language over a terminal alphabet Σ such that $|\Sigma|=1$. If $PL(L)$ holds, then L is a regular language.*

PROOF. Without loss of generality, let us consider a language L over the terminal alphabet $\{a\}$, such that $PL(L)$ holds. By Lemma 1, we have that $PL1(L)$ holds for some positive integer b. Let us consider the following two disjoint languages whose union is L:

(i) $L_{<b} = \{a^k \mid a^k \in L \wedge k<b\}$ and (ii) $L_{\geq b} = \{a^k \mid a^k \in L \wedge k\geq b\}$.

Now, $L_{<b}$ is a regular language, because it is finite. Since regular languages are closed under finite union and intersection [5], in order to prove that L is regular, it is enough to prove, as we now do, that

$$L_{\geq b} = \bigcup \mathcal{S}\ \cap\ \{a^i \mid i\geq b\} \qquad (\dagger 1)$$

where: (i) $\mathcal{S}$ is a set of languages which is a subset of the following finite set $\mathcal{L}$ of languages ($k, p_h, q_0, \ldots, q_k$ are integers):

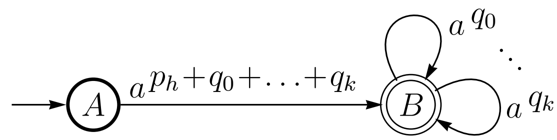
$$\begin{aligned}\mathcal{L} = \{L^{\langle p_h,\, q_0,\, \ldots,\, q_k\rangle} \mid\ & (0\leq k<b)\ \wedge\ (0\leq p_h<b)\ \wedge \\ & (0<q_0\leq b)\ \wedge \ldots \wedge\ (0<q_k\leq b)\ \wedge \\ & q_0, \ldots, q_k \text{ are all distinct}\,\}\end{aligned} \qquad (\dagger 2)$$

and (ii) for all $k, p_h, q_0, \ldots, q_k$, the language:

$$L^{\langle p_h, q_0, \ldots, q_k \rangle} = \{a^{p_h + i_0\, q_0 + \ldots + i_k\, q_k} \mid i_0 > 0 \;\wedge \ldots \wedge\; i_k > 0\} \qquad (\dagger 3)$$

is regular.

Indeed, (i) $\{a^i \mid i \geq b\}$ is regular, (ii) $\mathcal{L}$ is a finite set of languages because, for any b, there exists only a finite number of tuples $\langle p_h, q_0, \ldots, q_k\rangle$ satisfying all the conditions stated inside the set expression ($\dagger$2), and (iii) the language $L^{\langle p_h, q_0, \ldots, q_k \rangle}$ is regular because it is recognized by the following nondeterministic finite automaton with initial state A and final state B:



In order to prove Equality ($\dagger$1) it remains to prove that, for any $z \in L_{\geq b}$, there exists a tuple of the form $\langle p_h, q_0, q_1, \ldots, q_k\rangle$ such that $z \in L^{\langle p_h, q_0, q_1, \ldots, q_k \rangle}$.

Given any word $z \in L_{\geq b}$, the following algorithm constructs a tuple of the form $\langle p_h, q_0, q_1, \ldots, q_h\rangle$, for some $h \geq 0$.

$\{\, z \in L_{\geq b} \,\}$ *Tuple Generation Algorithm*

$\ell := |z|$; $\quad i := 0$; $\quad \langle p_0, q_0\rangle := \pi(\ell)$;

$\{\, |z| = p_i + \sum_{j=0}^{i} q_j \;\;\wedge\;\; \bigwedge_{j=0}^{i}\; 0 < q_j \leq b \;\;\wedge\;\; 0 \leq p_i \,\}$

while $p_i \geq b$ **do** $\ell := p_i$; $\;i := i+1$; $\;\langle p_i, q_i\rangle := \pi(\ell)$ **od**;

$h := i$;

$\{\, |z| = p_h + \sum_{j=0}^{h} q_j \;\;\wedge\;\; \bigwedge_{j=0}^{h}\; 0 < q_j \leq b \;\;\wedge\;\; 0 \leq p_h < b \,\}$

In this algorithm π is a function from N to $N \times N$, whose existence follows from the validity of $PL1(L)$, satisfying the following condition: for every $\ell \geq b$, $\pi(\ell) = \langle p, q\rangle$ such that $\ell = p + q$ and $0 < q \leq b$ (take $i = 1$ in Condition (4.1) of $PL1(L)$ in Lemma 1). The termination of the Tuple Generation Algorithm is a consequence of the fact that, for every $z \in L_{\geq b}$, for every $i \geq 0$, $p_i = p_{i+1} + q_{i+1}$ and $q_i > 0$. This implies that $p_0, p_1, \ldots$ is a strictly decreasing sequence of integers, and eventually in that sequence we will get an element smaller than b, and the while-loop terminates.

Thus, for every $z \in L_{\geq b}$, there exist $h \geq 0$, p_0, q_0, p_1, q_1, p_2, q_2, $\ldots$, p_h, q_h such that:

$$z = a^{(p_h + q_h) + q_{h-1} + \ldots + q_2 + q_1 + q_0} \qquad (\dagger 4)$$

where: $0 \leq p_h < b$ and for every i, if $0 \leq i < h$, then ($p_i \geq b$ and $0 < q_i \leq b$).

In general, in Equality (†4) the q_i's are *not* all distinct. Thus, by rearranging the summands, and writing $i_j\, q_j$, instead of $(q_j + \ldots + q_j)$ with i_j occurrences of q_j, we have that, for every word $z \in L_{\geq b}$, there exist some integers $k, p_h, i_0, q_0, \ldots, i_k, q_k$ such that

$z = a^{\,p_h \,+\, i_0\, q_0 \,+\, \ldots \,+\, i_k\, q_k}$, where:

$(\ell 0)$ $0 \leq k$, $(\ell 1)$ $0 \leq p_h < b$, $(\ell 2)$ $0 < q_0 \leq b, \ldots, 0 < q_k \leq b$,
$(\ell 3)$ $q_0, \ldots, q_k$ are all distinct, and $(\ell 4)$ $i_0 > 0, \ldots, i_k > 0$.

From $(\ell 2)$ and $(\ell 3)$, we have that $k < b$. Hence, Condition $(\ell 0)$ can be strengthened to: $(\ell 0^*)$ $0 \leq k < b$. We also have that $k \leq h$, and $k = h$ when in Equality (†4) the values of $q_0, \ldots, q_h$ are all distinct.

Since Conditions $(\ell 0^*)$, $(\ell 1)$, $(\ell 2)$, and $(\ell 3)$ are those occurring in the set expressions (†2), and Condition $(\ell 4)$ is the one occurring in the set expressions (†3), we have concluded the proof of Equality (†1) and that of Theorem 2. $\square$

Let us make a few of remarks on the proof of Theorem 2.

(i) The validity of $PL1(L)$ tells us that the function π exists, but it does not tell us how to compute $\pi(\ell)$, for any given $\ell \geq b$.

(ii) Since summation is commutative, it may be the case that a language in $\mathcal{L}$ corresponds to more than one tuple $\langle p_h, q_0, \ldots, q_k\rangle$. In particular, we have that $L^{\langle p_h,\, q_0,\, \ldots,\, q_k\rangle} = L^{\langle p_h,\, q'_0,\, \ldots,\, q'_k\rangle}$, whenever $\langle q_0, \ldots, q_k\rangle$ is a permutation of $\langle q'_0, \ldots, q'_k\rangle$.

(iii) If $b = 1$, then $k = h = 0$. Thus, from Conditions $(\ell 1)$ and $(\ell 3)$ we have: $\langle p_h, q_h\rangle = \langle p_0, q_0\rangle = \langle 0, 1\rangle$. We also have that $\mathcal{L}$ is the singleton $\{L^{\langle 0,\, 1\rangle}\}$, where $L^{\langle 0,\, 1\rangle}$ is the language $\{a^i \mid i > 0\}$.

(iv) In Equality (†1) the set $\mathcal{S}$ of languages may be a *proper* subset of $\mathcal{L}$. Indeed, let us consider the language $L = \{a\,(a\,a)^n \mid n \geq 0\}$ generated by the context-free grammar $S \to a\,S\,a \mid a$. Since $PL1(L)$ holds for 3, we can take the constant b occurring in Equality (†1) to be 3. If we consider the word $z = a\,a\,a$, then the set $\mathcal{L}$ of languages includes, among others, the languages $L^{\langle 0,3\rangle} = \{a^{0\,+\,i\cdot 3} \mid i > 0\}$, $L^{\langle 1,2\rangle} = \{a^{1\,+\,i\cdot 2} \mid i > 0\}$, and $L^{\langle 2,1\rangle} = \{a^{2\,+\,i\cdot 1} \mid i > 0\}$ (these three languages are obtained for $k = h = 0$). Now, $L_{\geq\, 3} = L^{\langle 1,2\rangle} \cap \{a^i \mid i \geq 3\} = L^{\langle 1,2\rangle}$, while $L^{\langle 0,3\rangle} \not\subseteq L_{\geq\, 3}$ and $L^{\langle 2,1\rangle} \not\subseteq L_{\geq\, 3}$.

(v) It may be the case that the length $p_h + q_0 + \ldots + q_k$ of the word labeling the arc from state A to state B of the finite automaton depicted above, is smaller than b. Thus, in the definition of $L_{\geq b}$ the intersection of $\bigcup \mathcal{S}$ with $\{a^i \mid i \geq b\}$ ensures that only words whose length is at least b are considered.

Acknowledgements

This work has been partially funded by INdAM-GNCS (Italy).

References

[1] Ş. Andrei, S. V. Cavadini, W.-N. Chin, A new algorithm for regularizing one-letter context-free grammars, Theoretical Computer Science 306 (1-3) (2003) 113–122.

[2] Y. Bar-Hillel, M. Perles, E. Shamir, On formal properties of simple phrase structure grammars, Z. Phonetik. Sprachwiss. Kommunikationsforsch. 14 (1961), 143–172.

[3] S. Ginsburg, G. H. Rice, Two families of languages related to ALGOL, JACM 9 (3) (1962), 350–371.

[4] M. A. Harrison, Introduction to Formal Language Theory, Addison Wesley, 1978.

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

[6] R. J. Parikh, On context-free languages, J. ACM 13 (4) (1966), 570–581.

[7] G. Ramos-Jiménez, J. López-Muñoz, R. Morales-Bueno, Comparisons of Parikh's condition to other conditions for context-free languages, Theoretical Computer Science 202 (1) (1998), 231 – 244.