Proceedings of the 8^{th} International Conference on Applied Informatics
Eger, Hungary, January 27–30, 2010. Vol. 2. pp. 73–81.

Pumping Lemmas for Special Linear Languages[*]

Benedek Nagy

Faculty of Informatics, University of Debrecen, Debrecen, Hungary
e-mail: nbenedek@inf.unideb.hu

The paper is dedicated to Pál Dömösi on his 65th birthday.

Abstract

Pumping lemmas give efficient way to prove that some languages do not belong to certain language classes. There are several known pumping lemmas mainly for context-free languages and some of its special classes. In this paper we present pumping lemmas for special linear context-free language classes where the rules of the grammar have strict restriction on their length. These lemmas can also be used in a non-standard way for regular languages pumping the words in two places simultaneously. We are presenting such kind of applications also.

Keywords: k-rated linear languages, pumping lemma, derivation tree, regular languages

MSC: 68Q45 Formal languages and automata

1. Introduction

The formal language theory and generative grammars form one of the basics of the field of theoretical computer science ([5]). Pumping lemmas play important role in formal language theory ([3, 4]). One can prove that a language does not belong to a given language class. There are well-known pumping lemmas, for example, for regular and context-free languages. The first and most known pumping lemma is introduced by Bar-Hillel, Perles, and Shamir in 1961 for context-free languages [3]. Nowadays several pumping lemmas are known for various language classes. Several subclasses of context-free languages are known, such as deterministic context-free and linear languages. The class of linear languages is strictly between the regular and the context-free ones. In linear grammars the following types of rules are

[*] The work is partially supported by NKTH, and a Japan-Hungarian bilateral project (TéT).

used: $A \to w$, $A \to uBv$ (A, B are non-terminals, $w, u, v \in V^*$). In the sixties, Amar and Putzolu defined and analysed a special subclass of linear languages, the so-called even-linear ones, in which the rules has a kind of symmetric shape [1]. The even-linear languages are intensively studied, for instance, they play special importance in learning theory [9]. In [2] Amar and Putzolu extended the definition to any fix-rated linear languages. They defined the k-rated linear grammars and languages, in which the ratio of the lengths of v and u equals to a fixed non-negative rational number k for all rules of the grammar containing non-terminal in the right-hand-side. They used the term k-linear for the grammar class and k-regular for the generated language class. In the literature the k-linear grammars and languages are frequently used for the metalinear grammars and languages [5], as they are extensions of the linear ones (having at most k nonterminals in the sentential forms). Therefore, for clarity, we prefer the term fix-rated (k-rated) linear for those restricted linear grammars and languages that are introduced in [2]. The classes of k-rated linear languages are strictly between the linear and regular ones for any positive rational value of k. Moreover their union, the set of all fixed-linear languages is also strictly included in the class of linear languages. Amar and Putzolu stated an open problem whether the intersection of the classes of the k-rated linear languages ($k > 0$) is exactly the class of the regular languages or not. The derivation-trees of the k-rated linear grammars form pine tree shapes. In this paper we investigate pumping lemmas for these languages. These new pumping lemmas work for regular languages as well, since every regular language is k-linear for every non-negative rational k. In this way the words of a regular language can be pumped in two places in a parallel way. It is an unusual way to pump regular languages since the previously known pumping lemmas pump the words in one place at a time.

2. Preliminaries

In this section we give some basic concepts and fix our notation. Let $\mathbb{N}$ denote the non-negative integers and $\mathbb{Q}$ denote the non-negative rational numbers through the paper. A grammar is a construct $G = (N, V, S, H)$, where N, V are the non-terminal and terminal alphabets. $S \in N$ is the initial letter. H is a finite set of derivation rules. A rule is a pair written in the form $v \to w$ with $v \in (N \cup V)^* N (N \cup V)^*$ and $w \in (N \cup V)^*$. Let G be a grammar and $v, w \in (N \cup V)^*$. Then $v \Rightarrow w$ is a direct derivation if and only if there exist $v_1, v_2, v', w' \in (N \cup V)^*$ such that $v = v_1 v' v_2$, $w = v_1 w' v_2$ and $v' \to w' \in H$. The transitive and reflexive closure of $\Rightarrow$ is denoted by $\Rightarrow^*$. The language generated by a grammar G is $L(G) = \{w | S \Rightarrow^* w \wedge w \in V^*\}$. Two grammars are equivalent if they generate the same language modulo the empty word (λ). (From now on we do not care whether $\lambda \in L$ or not.) Depending on the possible structures of the derivation rules we are interested in the following classes [2, 5].

• type 2, or context-free (CF) grammar: each rule has the following form: $A \to v$ with $A \in N$ and $v \in (N \cup V)^*$.

• linear (Lin) grammar: each rule has one of the next forms: $A \to v$, $A \to vBw$; where $A, B \in N$ and $v, w \in V^*$.
• k-rated linear (k-Lin) grammar: it is a linear grammar with the following property: there exists a rational number k such that for each rule of the form: $A \to vBw$: $\frac{|w|}{|v|} = k$ (where $|v|$ denotes the length of v).
• Specially with $k = 1$: even-linear (1-Lin) grammar.
• Specially with $k = 0$: type 3, or regular (Reg) grammar: each rule has one of the following forms: $A \to w$, $A \to wB$; where $A, B \in N$ and $w \in V^*$.
The language family regular/linear etc. contains all languages that can be generated by regular/linear etc. grammars. We call a language L fix-rated linear if there is a $k \in \mathbb{Q}$ such that L is k-rated linear. So the class of fix-rated linear languages includes all the k-rated linear language families. Moreover it is known [2], that for any $k \in \mathbb{Q}$ all regular languages are k-rated linear. Further, when we consider a special fixed value of k, then we will also use it as $k = \frac{g}{h}$, where $g, h \in \mathbb{N}$ $(h \neq 0)$ are relatively primes.
Now we present normal forms for the linear, k-rated linear and so, even-linear and regular grammars. The following fact is well-known:
Every linear grammar has an equivalent grammar in which all rules are in forms of $A \to aB, A \to Ba, A \to a$ with $a \in V, A, B \in N$.

Lemma 2.1 (Normal Form for ***k***-rated Linear Grammars). *Every k-rated (for $k = \frac{g}{h}$) linear grammar has an equivalent one in which for every rule of the form $A \to vBw$: $|w| = g$ and $|v| = h$ such that g and h are relatively primes and for all rules of the form $A \to u$ with $u \in V^*$: $|u| < g + h$ holds.*

The proof of this lemma goes in the standard way: longer rules can be simulated by shorter ones by the help of newly introduced new nonterminals.

As special cases of the previous lemma we have:
Every even-linear grammar has an equivalent grammar in which all rules are in forms $A \to aBb, A \to a$, $A \to \lambda$ $(A, B \in N, a, b \in V)$.
Every regular language can be generated by grammar having only rules of types $A \to aB, A \to \lambda$ $(A, B \in N, a \in V)$.

Derivation trees are widely used graphical representations of derivations in context-free grammars. The root of the tree is a node labelled by the initial symbol S. The terminal labelled nodes are leaves of the tree. The nonterminals, as the derivation continues from them, have some children nodes. In linear case, there is at most one non-terminal in every level of the tree. Therefore the derivation can go only in a linear (sequential) manner. There is only one main branch of the derivation (tree trunk); all the other branches terminate immediately. Observing the derivations and derivation trees for linear grammars, they seem to be highly related to the regular case. The linear (and so, specially, the even-linear and fixed linear) languages can be accepted by finite state machines [1, 7]. Moreover the k-rated linear languages are accepted by deterministic machines ([7, 8]). By an analysis of the possible trees and iterations of nonterminals in a derivation (tree) one can obtain pumping (or iteration) lemmas. Further in this section we recall some well-known pumping lemmas (see [3, 5]).

Lemma 2.2 (Bar-Hillel Lemma). *Let L be a context-free language. Then there exists an integer $n \in \mathbb{N}$ such that any word $p \in L$ with $|p| \geq n$, admits a factorization $p = uvwxy$ satisfying*
1. $uv^iwx^iy \in L$ for all $i \in \mathbb{N}$ 2. $|vx| > 0$ 3. $|vwx| \leq n$.

Lemma 2.3 (Pumping Lemma for Linear Languages). *Let L be a linear language. Then there exists an integer n such that any word $p \in L$ with $|p| \geq n$, admits a factorization $p = uvwxy$ satisfying*
1. $uv^iwx^iy \in L$ for all integer $i \in \mathbb{N}$ 2. $|vx| > 0$ 3. $|uvxy| \leq n$.

It can be shown that $\{a^ib^ic^jd^j | i, j \in \mathbb{N}\}$ is not linear. We note here that in [6] there is a pumping lemma for non-linear context-free languages.

Lemma 2.4 (Pumping Lemma for Regular Languages). *Let L be a regular language. Then there exists an integer n such that any word $p \in L$ with $|p| \geq n$, admits a factorization $p = uvw$ satisfying*
1. $uv^iw \in L$ for all integer $i \in \mathbb{N}$ 2. $|v| > 0$ 3. $|uv| \leq n$.

By the previous lemma one can easily show that $\{a^nb^n | n \in \mathbb{N}\}$ is not regular. In the next section we present pumping lemmas for the k-rated linear languages.

3. Main Results

Let us consider a k-rated linear grammar. Based on the normal form (Lemma 2.1) every word of a $k = \frac{g}{h}$-rated linear language can be generated by a 'pine-tree' shape derivation tree (i.e., the trunk has equal size branches on the right side and also equal size branches on the left side, see Fig. 1).

Now we are ready to present our new pumping lemmas.

Theorem 3.1. *Let L be a $(\frac{g}{h} = k)$-rated linear language. Then there exists an integer n such that any word $p \in L$ with $|p| \geq n$, admits a factorization $p = uvwxy$ satisfying*
1. $uv^iwx^iy \in L$ for all integer $i \in \mathbb{N}$ 2. $0 < |u|, |v| \leq n\frac{h}{g+h}$
3. $0 < |x|, |y| \leq n\frac{g}{g+h}$ 4. $\frac{|x|}{|v|} = \frac{|y|}{|u|} = \frac{g}{h} = k$.

Theorem 3.2. *Let L be a $(\frac{g}{h} = k)$-rated linear language. Then there exists an integer n such that any word $p \in L$ with $|p| \geq n$, admits a factorization $p = uvwxy$ satisfying*
1. $uv^iwx^iy \in L$ for all integer $i \in \mathbb{N}$ 2. $0 < |v| \leq n\frac{h}{g+h}$
3. $0 < |x| \leq n\frac{g}{g+h}$ 4. $0 < |w| \leq n$ 5. $\frac{|x|}{|v|} = \frac{|y|}{|u|} = \frac{g}{h} = k$.

The proofs of these theorems are based on the same idea as the proof of the Bar-Hillel lemma.

Remark 3.3. In case of $k = 0$ the previous theorems give the well-known pumping lemmas for regular languages.

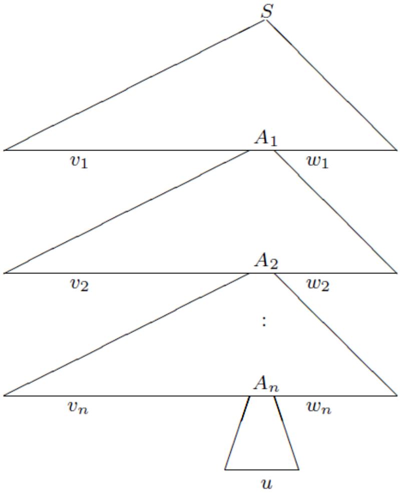

Figure 1: A 'pine-tree' shape derivation tree in a fix-rated linear grammar

4. Applications

As pumping lemmas are usually used to show that a language does not belong to a language class, we present an example for this type of application.

Example 4.1. The DYCK language (the language of correct bracket expressions) is not k-linear for any value of k over the alphabet $\{(,)\}$. Let $k \neq 1$ be fixed as $\frac{g}{h}$. Let us consider the word of the form $(^{(g+h)(n+2)})^{(g+h)(n+2)}$. Then Theorem 3.1 does not work (if $k \neq 1$), the pumping deletes or introduces different number of ('s and)'s. To show that the DYCK language is not 1-rated (i.e., even-)linear let us consider the word $(^{2n})^{2n}(^{2n})^{2n}$. Using Theorem 3.2 the number of inner brackets can be pumped. In this way such words are obtained in which there are prefixes with more letters) than (. Thus this language is not fixed-linear.

In the next example we consider a deterministic linear language.

Example 4.2. Let $L = \{a^m b^m | m \in \mathbb{N}\} \cup \{a^m c b^{2m} | m \in \mathbb{N}\}$. Let us assume that the language is fix-rated linear. First we show that this language is not fix-rated linear with ratio other than 1. Contrary, assume that it is, with $k = \frac{g}{h} \in \mathbb{Q}$ such that $k \neq 1$. Let n be given by Theorem 3.1. Then consider the words of the form $a^{m(g+h)} b^{m(g+h)}$ with $m > n$. By the theorem any of them can be factorized to $uvwxy$ such that $|uv| \leq \frac{2nh}{g+h}$. Since $g + h > 2$ (remember that $g, h \in \mathbb{N}$, relatively primes and $g \neq h$), $|uv| < nh$, and therefore both u and v contains only a's. By a similar argument on the length of xy, x and y contains only b's. Since the ratio $\frac{|x|}{|v|}$ (it is fixed by the theorem) is not 1, by pumping we get words outside of the language. Now we show that this language is not even-linear. Assume that it is 1-rated linear ($g = h = 1$). Let n be the value from Theorem 3.1. Let us consider the words of shape $a^m cb^{2m}$ with $m > n$. Now we can factorize these words in a way, that $|uv| \leq n$ and $|xy| \leq n$ and $|v| = |x|$. By pumping we get words $a^{m+j}cb^{2m+j}$ with some positive values of j, but they are not in L. Thus L is not fix-rated linear.

In the next example we show a fix rated linear language that can be pumped.

Example 4.3. Let L be the language of palindromes over $\{a, b\}$ (words p that are the same in reverse order ($p = p^R$)). Our pumping lemmas work with $k = 1$: Let $p \in L$, then $p = uvwxy$ according to Theorem 3.1 or Theorem 3.2, such that $|u| = |y|$ and $|v| = |x|$. By the property of the palindromes, we have $u = y^R$, $v = x^R$ and $w = w^R$. By $i = 0$ the word uwy is obtained which is in L by the previous equalities. By further pumping the words uv^iwx^iy are obtained, they are in L. Using Theorem 3.1 and considering words $a^m ba^m$, it can be shown in analogous way as in Example 4.2 that enough long words cannot be pumped with $k \neq 1$.

Besides our theorems work for regular languages with $k = 0$ there is a non-standard application of them. As we already mentioned, all regular languages are k-rated linear for any values of $k \in \mathbb{Q}$. Therefore every new pumping lemma works for any regular language with any values of k. Let us see an example.

Example 4.4. Let $L = (ab)^*aa(bbb)^*a$ be regular. We show that our theorems work for, let us say, $k = \frac{1}{2}$. Every word is of the form $(ab)^n aa(bbb)^m a$ (with $n, m \in \mathbb{N}$). For words that are long enough either n or m (or both of them) are sufficiently large. Now we detail effective factorizations $p = uvwxy$ of the possible cases. We give only those words of the factorization that have maximized lengths due to the applied theorem, the other words can easily be found by the factorization and, at Theorem 3.2, by taking into account the fixed ratio of some lengths.

- Theorem 3.1 for $k = \frac{1}{2}$:

if $n > 3$ and $m > 0$:	let $u = ab$,	$v = ababab$,	$x = bbb$,	$y = a$.
if $m = 0$:	let $u = ababab$,	$v = abab$,	$x = ab$,	$y = aaa$.
if $n = 3$:	let $u = abababaa$,	$v = bb$,	$x = b$,	$y = bbba$.
if $n = 2$:	let $u = ababaa$,	$v = bb$,	$x = b$,	$y = bba$.
if $n = 1$:	let $u = abaa$,	$v = bb$,	$x = b$,	$y = ba$.
if $n = 0$:	let $u = aa$,	$v = bb$,	$x = b$,	$y = a$.

- Theorem 3.2 for $k = \frac{1}{2}$:

if $n \leq 3m-4$:	let $v = bb$,	$w = b$	$x = b$.
if $n = 3m-3$:	let $v = ababab$,	$w = aabbbb$	$x = bbb$.
if $n = 3m-2$:	let $v = ababab$,	$w = abaabbbb$	$x = bbb$.
if $n = 3m-1$:	let $v = ababab$,	$w = ababaabbbb$	$x = bbb$.
if $n = 3m$:	let $v = ababab$,	$w = aab$	$x = bbb$.
if $n = 3m+1$:	let $v = ababab$,	$w = abaab$,	$x = bbb$.
if $n = 3m+2$:	let $v = ababab$,	$w = ababaab$,	$x = bbb$.
if $n = 3m+3$:	let $v = ababab$,	$w = abababaab$,	$x = bbb$.
if $n = 3m+4$:	let $v = ababab$,	$w = ababababaab$,	$x = bbb$.
if $n = 3m+5$:	let $v = ababab$,	$w = abababababaab$,	$x = bbb$.
if $n \geq 3m+6$, $n \equiv 0 (\text{mod} 3)$:	let $v = abab$,	$w = \lambda$,	$x = ab$.
if $n \geq 3m+7$, $n \equiv 1 (\text{mod} 3)$:	let $v = abab$,	$w = ab$,	$x = ab$.
if $n \geq 3m+8$, $n \equiv 2 (\text{mod} 3)$:	let $v = abab$,	$w = abab$,	$x = ab$.

In a similar way, it can be shown that pumping the words of a regular language in two places simultaneously with other values of k (for instance, $1, 5, \frac{7}{3}$ etc.) works. In the next example we show that there are languages that can be pumped by the usual pumping lemmas for regular languages, but they cannot be regular since we prove that there is a value of k such that one of our theorems does not work.

Example 4.5. Let $L = \{a^r b a^q b^m | r, q, m \geq 2, \exists j \in \mathbb{N} : q = j^2\}$. By the usual pumping lemmas for regular languages, i.e., by fixing k as 0, one cannot infer that this language is not regular. By $k = 0$, $x = y = \lambda$ and so $p = uvw$. Due to the a's in the beginning, Theorem 3.1 works: $u = a, v = a$; and due to the b's in the end, Theorem 3.2 also works: $v = b, w = b$. Now we show that L is not even-linear. Contrary, let us assume that Theorem 3.2 works for $k = 1$. Let n be the value for this language according to the theorem. Let $p = a^2 b a^{(2n+5)^2} b^3$. By the conditions of the theorem, it can be factorized to $uvwxy$ such that $|v|, |w|, |x| \leq n$ and $|u| = |y|$. In this way vwx must be a subword of $a^{(2n+5)^2}$, and so, the pumping decrease/increase only q. Since $|v|, |x| \leq n$ in the first round of pumping $p' = a^2 b a^{(2n+5)^2 + |vx|} b^3$ is obtained. But $(2n+5)^2 < (2n+5)^2 + |vx| \leq (2n+5)^2 + 2n < (2n+6)^2$, therefore $p' \notin L$.
Thus L is not even-linear, and therefore it cannot be regular. Our pumping lemma was effective to show this fact.

Usually pumping lemmas can be used only to show that some languages do not belong to the given class of languages. One may ask what we can say if a language satisfies our theorems. We cannot infer about the language class if a language satisfies our new pumping lemmas:

Example 4.6. Let $L = \{0^j 1^m 0^r 1^i 0^l 1^i 0^r 1^m 0^j | j, m, i, l, r \geq 1, r \text{ is prime}\}$. One can easily show that this language satisfies both Theorem 3.1 and Theorem 3.2 with $k = 1$: one can find subwords to pump in the part of outer 0's or 1's (pumping their number from a given j or m to arbitrary high values), or in the middle part 0's or 1's (pumping their number from i or l to arbitrary high values), respectively. But this language is not even context-free, since intersected by the regular language 010^*1010^*10 a non semi-linear language is obtained. Since context-free languages are semi-linear (Parikh theorem) and the class of context-free languages are closed under intersection with regular languages, we just proved that L is not linear and so, not fix-rated linear.

It is a more interesting question what we can say about a language for which there are values $k_1 \neq k_2$ such that all its enough long words can be pumped both as k_1-rated and k_2-rated linear language. We hope our results can be used to solve the long-standing open problem of Amar and Putzolu [2], i.e., the problem if the intersection of the classes of k-rated linear languages ($k > 0$) is exactly the class of regular languages. We have the following conjecture.

Conjecture 4.7. If a language L satisfies any of our pumping lemmas for two different values of k, then L is regular.

If the previous conjecture is true, then exactly the regular languages form the intersection of the k-rated linear language families (for $k \in \mathbb{Q}$).

5. Conclusions

In this paper some pumping lemmas are proved for special linear languages. In fix-rated linear languages the lengths of the pumped subwords of a word depend on each other, therefore these pumping lemmas are more restricted than the ones working on every linear or every context-free languages. Since all regular languages are k-rated linear for any $k \in \mathbb{Q}$, these lemmas also work for regular languages. The question whether only regular languages satisfy our pumping lemmas for at least two different values of k (or for all values of k) is remained open as a conjecture. If the conjecture was found to be true, then the open problem of Amar and Putzolu will be solved affirmatively.

References

[1] Amar, V. and Putzolu, G.R.: On a Family of Linear Grammars, *Information and Control* **7**/3 (1964), 283–291.

[2] Amar, V. and Putzolu, G.R.: Generalizations of Regular Events, *Information and Control* **8**/1 (1965), 56–63.

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

[4] Dömösi, P., Ito, M., Katsura, M. and Nehaniv, C.: New Pumping Lemma for Context-Free Languages, *Proceedings of DMTCS'96*, 187–193.

[5] Hopcroft, J. E., Ullman, J. D., Introduction to Automata Theory, languages, and Computation, *Addison-Wesley, Reading, Mass.,* (1979).

[6] Horváth, G.: New Pumping Lemma for Non-Linear Context-Free Languages, *Proc. 9th Symp. Algebras, Lang. and Computation*, Shimane University, (2006), 160–163.

[7] Nagy, B.: On $5' \to 3'$ sensing Watson-Crick finite automata, *Proc. of DNA 13, LNCS* **4848** (2008), 256–262.

[8] Nagy, B.: On a hierarchy of $5' \to 3'$ sensing WK automata languages, *Abstract Booklet of Computability in Europe CiE 2009*, Heidelberg, (2009), 266–275.

[9] Sempere, J.M. and García, P.: A Characterization of Even Linear Languages and its Application to the Learning Problem, *Proc. ICGI-94, LNAI* **862** (1994), 38–44.

Benedek Nagy
University of Debrecen, Faculty of Informatics
4010 Debrecen, PO Box. 12, Hungary