International Journal of Mathematics and
Computer Science, **15**(2020), no. 4, 1173–1177

Language of Lyndon Partial Words

R. Krishna Kumari[1], R. Arulprakasam[2], V. R. Dare[3]

[1,2]Department of Mathematics
College of Engineering and Technology
SRM Institute of Science and Technology
SRM Nagar, Kattankulathur, 603203
Chennai, Tamilnadu, India

[3]Department of Mathematics
Madras Christian College
Chennai-600 059, Tamilnadu, India

email: kr1062@srmist.edu.in, r.aruljeeva@gmail.com,
rajkumardare@yahoo.com

(Received July 9, 2020, Accepted September 7, 2020)

Abstract

Combinatorial and algorithmic properties of Lyndon words have been investigated. Lyndon words have wide usage in algebra and pattern matching. In this paper, we introduce Lyndon partial words and show that the language of all Lyndon partial words over the binary alphabet is not context free.

1 Introduction

In 1954 Roger Lyndon introduced Lyndon words under the name of standard lexicographic sequences. Lyndon words are also widely used for computing the shortest superstring for a set of strings, bases construction in free Lie Algebras [7], de Brujin sequences construction, musicology, bio-informatics, pattern matching to name a few. Duval [9] presented an efficient algorithm

Key words: Context free language, Lyndon words, Partial words
AMS (MOS) Subject Classifications: 68Q45, 68Q70
ISSN 1814-0432, 2020, http://ijmcs.future-in-tech.net

to obtain a factorization of words over an ordered alphabet known as Lyndon factorization. Siromaney et al. [11] defined infinite Lyndon words and discussed their combinatorial and topological properties. Partial words are currently analyzed due to their usage in DNA computations. In DNA sequencing, some unseen or missing part of information may exist. This can be revealed by positions denoting don't care symbols in a word. Thus, instead of complete words, partial words are considered. In 1974, Fisher and Parterson [8] introduced partial words as strings with don't care symbols. In [1] Berstel and Boasson initiated "the study of combinatorics on partial words". This study was later pursued by Blanchet-Sadri et al. [3, 4, 5, 6]. Motivated by the application of Lyndon words and partial words in pattern matching and also by the work on [2, 11, 9, 10], the present paper extends Lyndon words to Lyndon partial words. Here we study properties of Lyndon partial words and show that language of all Lyndon partial words over the binary alphabet is not context free. The paper is organized as follows:
The basic notions used in this paper are given in Section 2. In Section 3, the language of Lyndon partial words is discussed.

2 Preliminaries

Here we recollect basic definitions and notations. Let Σ be a non-empty finite set of symbols. These symbols are called letters and the set is called an alphabet. Any string over Σ is called a *word*. The word p is a *subword* $(or factor)$ of q if there exists the words x and y such that $q = xpy$. If $xy \neq \epsilon$, then p is a proper subword of q. If $x = \epsilon$, then p is a prefix of q. If $y = \epsilon$, then p is a suffix of q. The sequence or word that contains a number of "do not know "symbols or "holes"denoted as $\Diamond$ are called *partial word*. The symbol $\Diamond$ does not belong to the alphabet Σ but a standby symbol for the unknown letter. A partial word of length n over $\Sigma_\Diamond$ is a partial function $r_\Diamond : \{0, 1, 2 \cdots, n-1\} \to \Sigma_\Diamond = \Sigma \cup \{\Diamond\}$ defined by $r_\Diamond(i) = r(i)$ if $i \in$ D(r) , $\Diamond$ if $i \in$ H(r), where $D(r)$ and $H(r)$ are the domain set and hole set of r respectively. A partial word x is said to be *primitive* if $\exists$ no word y such that $x = y^i$ with $i \geq 2$. A $finite\ Lyndon\ word\ l$ is a primitive word which is non-empty and less than all its conjugates in the alphabetical order(lexicographical order).

3 Main Results

Definition 3.1. *A finite Lyndon partial word $l_\Diamond$ is a primitive partial word which is non-empty and less than all its conjugates in the alphabetical order. The language $L_\Diamond$ represents the finite Lyndon partial words set over $\Sigma_\Diamond$. Equivalently, $l_\Diamond \in L_\Diamond$ iff $\forall t_\Diamond s_\Diamond \in \Sigma_\Diamond^+, l_\Diamond = t_\Diamond s_\Diamond \Rightarrow l_\Diamond < s_\Diamond t_\Diamond$.*

Example 3.1. Consider the ordered alphabet $\Sigma = \{a, b\}$ with order $\{a < b\}$. Let $L_\Diamond$ represent a finite Lyndon partial words set over $\Sigma_\Diamond = \{a, b\} \cup \{\Diamond\}$, where the symbol $\Diamond$ does not belong to the ordered alphabet Σ but a standby symbol for the unknown letter. The finite Lyndon partial words set with one hole and of length at most three are

$$L_\Diamond = \{a\Diamond, \Diamond b, a\Diamond b\}.$$

$\Diamond$ alone cannot be considered as a finite Lyndon partial word of length one since partial words are of at least length 2.

Theorem 3.1. *No proper subword exists as both prefix and suffix of a Lyndon partial word.*

Proof. Let $u_\Diamond$ be a partial word over $\Sigma_\Diamond^+$. Let $p_\Diamond$ be a proper subword of $u_\Diamond$ such that $p_\Diamond$ is both prefix and suffix of $u_\Diamond$. The partial word $u_\Diamond = p_\Diamond q_\Diamond^i$ and $u_\Diamond = q_\Diamond^j p_\Diamond$ for some $q_\Diamond^i, q_\Diamond^j \in \Sigma_\Diamond^+$. Now let us consider $u_\Diamond$ over $\Sigma_\Diamond^+$ belongs to $L_\Diamond$. By the notion of finite Lyndon words, $u_\Diamond < q_\Diamond^i p_\Diamond$ and $u_\Diamond < p_\Diamond q_\Diamond^j$. Then $q_\Diamond^j p_\Diamond < q_\Diamond^i p_\Diamond$ and $p_\Diamond q_\Diamond^i < p_\Diamond q_\Diamond^j$. This shows that $q_\Diamond^j < q_\Diamond^i$ and $q_\Diamond^i < q_\Diamond^j$ which is impossible. Therefore, $u_\Diamond$ over $\Sigma_\Diamond^+$ does not belongs to $L_\Diamond$. □

Theorem 3.2. *A partial word $u_\Diamond$ over $\Sigma_\Diamond^+$ belongs to $L_\Diamond$ iff $u_\Diamond < q_\Diamond$ for each proper suffix $q_\Diamond$ of $u_\Diamond$.*

Proof. Consider $u_\Diamond = p_\Diamond q_\Diamond$ to be a finite Lyndon partial word. Let $u_\Diamond^i \in L_\Diamond$ be the prefixes of $u_\Diamond$ such that $|p_\Diamond| < |u_\Diamond^i|$ for $i \in N$. This shows that $u_\Diamond^i = p_\Diamond q_\Diamond^i$, where $q_\Diamond^i \in \Sigma_\Diamond^+$. All $u_\Diamond^i$ are Lyndon and $u_\Diamond^i < q_\Diamond^i < q_\Diamond$. Thus $u_\Diamond < q_\Diamond$. Conversely, if $u_\Diamond < q_\Diamond$ for all suffix $q_\Diamond$ of $u_\Diamond$, then the prefixes $u_\Diamond^i$ of $u_\Diamond$ will be less than $y_\Diamond^j$, for all suffix $q_\Diamond^j$ of $u_\Diamond^i$. Thus $u_\Diamond^i$ is Lyndon. This implies that $u_\Diamond$ is a finite Lyndon partial word. □

Theorem 3.3. *Consider $p_\Diamond \in L_\Diamond$ and $q_\Diamond \in L_\Diamond$. Then $p_\Diamond q_\Diamond \in L_\Diamond$ iff $p_\Diamond < q_\Diamond$.*

Proof. Consider $p_\Diamond q_\Diamond \in L_\Diamond$. Then by Theorem 3.2, $p_\Diamond < p_\Diamond q_\Diamond < q_\Diamond$.
Conversely, consider $p_\Diamond < q_\Diamond$. Let r be a proper suffix of $p_\Diamond$ such that either

$r = q_\Diamond$ or
$Case(i)$: For $p^i_\Diamond$ a proper suffix of $p_\Diamond \in L_\Diamond$, $p_\Diamond < p^i_\Diamond$. Then $p_\Diamond q_\Diamond < p^i_\Diamond q_\Diamond$. If $p_\Diamond$ is not a prefix of $q_\Diamond$, then $p_\Diamond q_\Diamond < q_\Diamond$ since $p_\Diamond < q_\Diamond$. If $p_\Diamond$ is a prefix of $q_\Diamond$, then $y_\Diamond = p_\Diamond < q^j_\Diamond$ where $q^j_\Diamond$ is a proper suffix of $q_\Diamond$. Thus $q_\Diamond < q^j_\Diamond$ and $p_\Diamond q_\Diamond < p_\Diamond q^j_\Diamond$. Then we arrive at the result $p_\Diamond q_\Diamond < q_\Diamond$.
$Case(ii)$:For $q^i_\Diamond$ a proper suffix of $q_\Diamond \in L_\Diamond$, $q_\Diamond < q^i_\Diamond$. Then $p_\Diamond q_\Diamond < q_\Diamond < qi_\Diamond$. In both cases, we have $p_\Diamond q_\Diamond < r$. Therefore, by Theorem 3.2, we conclude that $p_\Diamond q_\Diamond \in L_\Diamond$. □

Theorem 3.4. *Any partial word $u_\Diamond$ over the alphabet $\Sigma^+_\Diamond$ can be uniquely written as $u_\Diamond = l^1_\Diamond.....l^r_\Diamond$ with $l^1_\Diamond, ..., l^r_\Diamond \in L_\Diamond$ and $l^1_\Diamond \geq ..... \geq l^r_\Diamond$.*

Proof. We have to show that any partial word resolves uniquely as a non-increasing product of Lyndon partial words. Since the symbols are in the finite set of Lyndon words $L_\Diamond$, any partial word has a resolution of factors in Lyndon partial words. Now, consider a resolution of factors $u_\Diamond = l^1_\Diamond.....l^r_\Diamond$ with r minimal. If $l^i_\Diamond < l^{i+1}_\Diamond$ for some i, then $u_\Diamond = l^1_\Diamond....l^{i-1}_\Diamond(l^i_\Diamond l^{i+1}_\Diamond)....l^r_\Diamond$ is a resolution of factors in Lyndon partial words since we have $l^i_\Diamond l^{i+1}_\Diamond \in L_\Diamond$. Now, we have to prove the uniqueness. Let us assume that for any $l^i_\Diamond, k^i_\Diamond \in L_\Diamond$ such that $l^1_\Diamond.....l^r_\Diamond = k^1_\Diamond.....k^r_\Diamond$, we have $l^1_\Diamond \geq ..... \geq l^r_\Diamond$ and $k^1_\Diamond \geq ..... \geq k^r_\Diamond$. Assume that $l^1_\Diamond$ is longer than $k^1_\Diamond$. Then $l^1_\Diamond = k^1_\Diamond.....k^i_\Diamond x$ with x a non empty prefix of $k^{i+1}_\Diamond$. Then $l^1_\Diamond < x \neq k^{i+1}_\Diamond \neq k^1_\Diamond l^1_\Diamond$ which contradicts our assumption. □

In formal language theory, a generalized version of pumping lemma known as Ogden's lemma [10] is used in cases where the pumping lemma is not sufficient to prove that certain languages are not context free.

Theorem 3.5. *$L_\Diamond$ of all Lyndon partial words over the alphabet $\Sigma^+_\Diamond$ is not context free.*

Proof. Assume that $L_\Diamond \subseteq \Sigma^+_\Diamond$ of Lyndon partial words is context free. Consider the word $l = a^{p+1}(\Diamond ba^p)^2b^2$. Let $l = uvwxy$ be a factorization and $p \geq 1$ be any integer as in Odgen's lemma. Then either v factor or x factor or both the factors are contained in the mid group.
$case(i)$: Both the factors u and v are in the mid group. Then iterating up we get a word of the form $l = a^{p+1}\Diamond ba^q \Diamond ba^p b^2$ with $q > p + 1$ which is not a Lyndon partial word since its conjugate $a^q \Diamond ba^p b^2 a^{p+1} \Diamond b \leq a^{p+1} \Diamond ba^q \Diamond ba^p b^2$.
$case(ii)$: The factor u is in the initial group and the factor v is in the secondary group of $a's$. Then iterating down we get a word of the form $a^r \Diamond ba^q \Diamond ba^p b^2$ where $r \leq p$ and $q < p$ which is not a Lyndon partial word since its conjugate $a^p b^2 a^r \Diamond ba^q \Diamond b \leq a^r \Diamond ba^q \Diamond ba^p b^2$.

case(iii) : The factor u is in the mid group and the factor v is in the final group of $a's$. Then iterating up we get a word of the form $a^{p+1}\Diamond ba^q\Diamond ba^r b^2$, where $r, q \geq p+2$ which is not a Lyndon partial word since its conjugate $a^r b^2 a^{p+1}\Diamond ba^q\Diamond b \leq a^{p+1}\Diamond ba^q\Diamond ba^r b^2$.

□

References

[1] J. Berstel, L. Boasson, Partial Words and a Theorem of Fine and Wilf, Theoret. Comput. Sci., **218**, (1999), 135–141.

[2] J. Berstel, L. Boasson, The language of Lyndon words is not context-free, Bull. EATCS, **63**, (1997).

[3] F. Blanchet-Sadri, A Periodicity Result of Partial Words with One Hole, Computers and Mathematics with Applications **46**, (2003), 813–820.

[4] F. Blanchet-Sadri, Periodicity on Partial Words, Computers and Mathematics with Applications **47,** (2004), 71–82.

[5] F. Blanchet-Sadri, Primitive Partial Words,Discrete Applied Mathematics **148**, (2005), 195–213.

[6] F. Blanchet-Sadri, R. A. Hegstrom, Partial Words and a Theorem of Fine and Wilf Revisited,*T*heoret. Comput. Sci. **270**, (2002), 401–419.

[7] Christophe Reutenauer, Free Lie Algebras, London Mathematical Society Monographs, New Series, **7**, (1993), The Clarendon Press, Oxford University Press, New York.

[8] M. J. Fischer, M. S. Paterson, String Matching and other Products (R. M. Karp, ed.), Complexity of Computation, SIAM-AMS Proceedings, (1974), 113–125.

[9] Jean-Pierre Duval, Factorizing Words over an Ordered Alphabet, J. Algorithms, **4,** no. 4, (1983), 363–381.

[10] Pal Domosi, Masami Ito, Context Free Languages and Primitive Words, World Scientific, (2014).

[11] Rani Siromoney, Lisa Mathew, V. R. Dare, K. G. Subramanian, Infinite Lyndon Words, Inform. Process. Lett., **50,** no. 2, (1994), 101–104.