CDMTCS
Research
Report
Series

On a problem of L. Priese

Ioan Tomescu
Faculty of Mathematics
University of Bucharest

CDMTCS-022
November 1996

Centre for Discrete Mathematics and
Theoretical Computer Science

On a problem of L. Priese *

Ioan Tomescu
Faculty of Mathematics,
University of Bucharest,
Str. Academiei, 14,
R-70109 Bucharest, Romania

Abstract

Lutz Priese raised the following conjecture: Almost all words of length n over a finite alphabet A with m letters contain as subwords all words of length $\lfloor loglogn \rfloor$ over A as $n \to \infty$. In this note we prove that this property holds for subwords of length $k(n)$ over A provided $lim_{n\to\infty} k(n)/logn = 0$.

1 Notation and preliminary results

Let A be a finite alphabet of cardinality $\mid A \mid= m$. A word $b \in A^*$ is said to be a subword of $a \in A^*$ if there exist $p, q \in A^*$ such that $a = pbq$. If $\alpha = \alpha_1 \dots \alpha_1 \in A^*$ is a fixed word of length $\mid \alpha \mid= k \geq 1$ having all letters equal to α_1, let $L(n)$ denote the number of words $a \in A^*$ such that $\mid a \mid= n$ and a does not contain the subword α.

Lemma 1.1 *We have*

$$L(n+1) = mL(n) - (m-1)L(n-k) \tag{1}$$

for every $n \geq k+1$ *and* $L(p) = m^p$ *for* $0 \leq p \leq k-1$ *and* $L(k) = m^k - 1$. *Moreover, the number of words* $a \in A^*$ *such that* $\mid a \mid= n$ *and* a *does not contain a fixed subword* $\beta = \beta_1 \dots \beta_k \neq \alpha$ *of length* k *over* A *is less than or equal to* $L(n)$.

Proof: It is clear that $L(p) = m^p$ for every $1 \leq p \leq k-1$, $L(k) = m^k - 1$ and $L(0) = 1$ since $L(k+1) = m^{k+1} - 2m + 1 = mL(k) - (m-1)L(0)$ by (1). In order to prove (1) let us denote by $W(n, k)$ the number of words w of length n over A not containing α such that $w = p\alpha_1 \dots \alpha_1$, where p is a word of length $n-k+1$ over A. It follows that $L(n+1) = m(L(n) - W(n,k)) + (m-1)W(n,k) = mL(n) - W(n,k)$ since the set of words of length $n+1$ over A not containing α can be generated in the following way: For every word a of length n over A not containing α we add a suffix consisting of any of m letters of A if $a \neq p\alpha_1 \dots \alpha_1$ ($\mid p \mid= n-k+1$) or a suffix consisting of any letter of A which is different from

*This work has been partially done while the author visited the Computer Science Department, University of Auckland, New Zealand

α_1 otherwise.
Now $W(n,k) = (m-1)L(n-k)$ since for every word $w = p\alpha_1 \dots \alpha_1$ not containing α and $| p |= n-k+1$, the prefix p can be obtained from a word u such that $| u |= n-k$ and u does not contain α by adding (in $m-1$ ways) a suffix consisting of one letter different from α_1 and (1) follows.
By denoting by $\mathcal{L}(n,\alpha)$ and $\mathcal{L}(n,\beta)$ the sets of words of length n over A containing a subword equal to α and to β, respectively, we will show that there exists an injective function φ: $\mathcal{L}(n,\alpha) \to \mathcal{L}(n,\beta)$. Indeed, let $a \in \mathcal{L}(n,\alpha)$ and denote by p, q and r ($| q |\geq k$) three subwords of a such that $a = pqr$; $q = \alpha_1 \dots \alpha_1$; $p, r \in A^*$; p does not contain a subword equal to α; the last letter of p and the first letter of r are different from α_1. For every such word $a \in\mathcal{L}(n,\alpha)$ we define $\delta(a) =| p |$ and let $\mathcal{M}(n,\alpha,\beta) \subset \mathcal{L}(n,\alpha)\cap \mathcal{L}(n,\beta)$ defined by

$$\mathcal{M}(n,\alpha,\beta) = \{a \in \mathcal{L}(n,\alpha) : a = u\beta v \text{ and } | u |\geq \delta(a)+1\},$$

i.e., the subset of words in $\mathcal{L}(n,\alpha)$ containing β as a subword such that the rank of the first letter of β is greater than the rank of the first letter of α. We define φ as follows:

$\varphi(a) = a$ if $a \in\mathcal{M}(n,\alpha,\beta)$;
$\varphi(a) = psr$, where $s = \beta\alpha_1 \dots \alpha_1$ is obtained by replacing in q the first k occurrences of α_1 by the word β if $a \notin\mathcal{M}(n,\alpha,\beta)$.

¿From this definition it follows that $\varphi(a) \in\mathcal{L}(n,\beta)\setminus \mathcal{M}(n,\alpha,\beta)$ whenever $a \in\mathcal{L}(n,\alpha)\setminus\mathcal{M}(n,\alpha,\beta)$ and φ is an injective function, which implies $|\mathcal{L}(n,\alpha) |\leq| \mathcal{L}(n,\beta) |$. It follows that $m^n - |\mathcal{L}(n,\beta) |\leq m^n - |\mathcal{L} (n,\alpha) |$ and the second assertion is verified.

□

Lemma 1.2 *For every $m, k \geq 2$ the characteristic equation of the recurrence (1),*

$$x^{k+1} - mx^k + m - 1 = 0 \tag{2}$$

has the following properties:
i) for k even it has roots $x_1, x_2, \dots, x_{k+1}$ such that $x_1 = 1$, $x_2 > 0$, $-1 < x_3 < -\frac{1}{2}$ and $x_4, \dots, x_{k+1}$ are complex numbers;
ii) for k odd it has $x_1 = 1$,$x_2 > 0$ and $x_3, \dots, x_{k+1}$ are complex numbers;
iii) all roots are simple;
iv) $m - \frac{1}{m^{k-1}} < x_2 < m - \frac{1}{m^k}$;
v) if z is a complex root of (2) then $| z |\leq 1$.

Proof: If $P(x) = x^{k+1} - mx^k + m - 1$ then $P(1) = 0$ and $P'(x)$ has a multiple root equal to zero and a simple root equal to $mk/(k+1)$; now i), ii) and iii) follow from the properties of the Rolle sequence associated to $P(x)$ and we obtain also $\frac{mk}{k+1} < x_2 < m$. We need a more accurate estimation for x_2, given by iv).
We get $P(m - \frac{1}{m^{k-1}}) < 0$ since this is equivalent to

$$m^{k^2-1}(m-1) < (m^k - 1)^k \tag{3}$$

By considering the function

$$f(x) = \frac{(m^x - 1)^x}{m^{x^2} - m^{x^2-1}},$$

we deduce that $f'(x) < 0$ for every $m \geq 2$ and $x > 1$ and $lim_{x\to\infty} f(x) = \frac{m}{m-1} > 1$, which proves (3). On the other hand $P(m - \frac{1}{m^k}) > 0$

because this is equivalent to $m^{k^2+k}(m-1) > (m^{k+1}-1)^k$. But this inequality holds since $m^{k^2+k} > (m^{k+1}-1)^k$.
In order to prove v) let z be a complex root of (2), $z = \rho(\cos\theta + i\sin\theta)$ $(\rho = \mid z \mid)$. From (2) it follows that $\rho^k \mid z - m \mid = m - 1$. But $\mid z - m \mid = \sqrt{\rho^2 - 2m\rho\cos\theta + m^2} \geq \mid \rho - m \mid$ and equality holds if and only if $\cos\theta = 1$, or $z \in \mathbb{R}$, which contradicts our hypothesis. It follows that

$$m - 1 > \rho^k \mid \rho - m \mid \tag{4}$$

By dividing $P(x)$ by $x-1$ we get $x^k - (m-1)(x^{k-1} + x^{k-2} + \ldots + 1)$, hence z is a root of the equation

$$x^k - (m-1)(x^{k-1} + x^{k-2} + \ldots + 1) = 0 \tag{5}$$

Suppose that $\rho > 1$. From (5) we deduce that

$$\rho^k \leq (m-1)(\rho^{k-1} + \ldots + 1) = (m-1)\frac{\rho^k - 1}{\rho - 1},$$

which is equivalent to

$$\rho^k(m - \rho) \geq m - 1 \tag{6}$$

¿From (6) we obtain $\rho < m$ and (4) becomes $m - 1 > \rho^k(m-\rho)$, which contradicts (6). It follows that $\rho \leq 1$.

□

Lemma 1.3 *The solution of the recurrence (1) is* $L(n) = \sum_{i=2}^{k+1} C_i x_i^n$*, where*

$$C_i = \frac{x_i(x_i - 1)}{(m-1)((k+1)x_i - km)}$$

for every $i = 2, \ldots, k+1$*.*

Proof: The solution of the recurrence (1) is of the form $L(n) = \sum_{i=1}^{k+1} C_i x_i^n$, where $x_1, \ldots, x_{k+1}$ are the solutions of (2) denoted as in Lemma 1.2. From the initial conditions given in Lemma 1.1 we get the following system of linear equations in $C_1, \ldots, C_{k+1}$:

$$\begin{array}{l} C_1 + C_2 + \ldots + C_{k+1} = 1 \\ C_1 x_1 + C_2 x_2 + \ldots + C_{k+1} x_{k+1} = m \\ \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \quad\quad \cdot \quad\quad \cdot \\ C_1 x_1^{k-1} + C_2 x_2^{k-1} + \ldots C_{k+1} x_{k+1}^{k-1} = m^{k-1} \\ C_1 x_1^k + C_2 x_2^k + \ldots + C_{k+1} x_{k+1}^k = m^k - 1 \end{array}$$

The determinant of the coefficient matrix is a Vandermonde determinant which will be denoted by $V(x_1, \ldots, x_{k+1}) = \prod_{1 \leq i < j \leq k+1} (x_j - x_i)$. Let us denote

$$S_M^{(p)}(x_1, \ldots, x_{k+1}) = \sum_{\substack{\{j_1, \ldots, j_p\} \subseteq M \\ 1 \leq j_1 < j_2 < \ldots < j_p \leq k+1}} x_{j_1} \ldots x_{j_p},$$

where $M \subseteq \{1,\ldots,k+1\}$ for $1 \leq p \leq \mid M \mid$ and $S_M^{(0)}(x_1,\ldots,x_{k+1}) = 1$ and consider the following matrix:

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^k & x_2^k & \cdots & x_{k+1}^k \end{pmatrix}$$

In order to compute some pseudo-Vandermonde determinants we recall the following result (see e.g. [4], problem 246): The determinant of the matrix deduced from this matrix by deleting the j-th row and the i-th column equals

$$S^{(k+1-j)}_{\{1,\ldots,i-1,i+1,\ldots,k+1\}}(x_1,\ldots,x_{k+1})V(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_{k+1})$$

Then by Cramer's rule we deduce that
$C_i = (-1)^{i-1}V(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_{k+1})(S_M^{(k)}(x_1,\ldots,x_{k+1}) - mS_M^{(k-1)}(x_1,\ldots,x_{k+1}) + m^2S_M^{(k-2)}(x_1,\ldots,x_{k+1}) - \ldots + (-1)^{k-1}m^{k-1}S_M^{(1)}(x_1,\ldots,x_{k+1}) + (-1)^k m^k + (-1)^{k+1})/V(x_1,\ldots,x_{k+1})$,
where $M = \{1,\ldots,i-1,i+1,\ldots,k+1\}$ has cardinality k.
Hence $C_i = (-1)^{i-1}V(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_{k+1})(-1)^k((m-x_1)\ldots(m-x_{i-1})(m-x_{i+1})\ldots(m-x_{k+1})-1)/V(x_1,\ldots,x_{k+1}) = (\frac{P(m)}{m-x_i}-1)/\prod_{\substack{j=1\\ j\neq i}}^{k+1}(x_i-x_j)$, where $P(x) = x^{k+1} - mx^k + m - 1$. We have $\prod_{\substack{j=1\\ j\neq i}}^{k+1}(x_i - x_j) = P'(x_i) = x_i^{k-1}((k+1)x_i - km)$, which implies

$$C_i = \frac{x_i - 1}{(m-x_i)x_i^{k-1}((k+1)x_i - km)}$$

Since $P(x_i) = 0$ we deduce that $(m-x_i)x_i^{k-1} = (m-1)/x_i$ and the formula for C_i follows. Note that $x_1 = 1$ implies $C_1 = 0$. □

2 Main result

Now we are ready to prove Priese's conjecture [3]. Let $\mathcal{W}(n,k,A)$ denote the set of words w of length n over A having the property that each word of length k over A is a subword of w.

Theorem 2.1 *If $lim_{n\to\infty}\frac{k(n)}{logn} = 0$, then almost all words of length n over A contain as subwords all words of length $k(n)$ over A as $n \to \infty$, i.e., $lim_{n\to\infty}\frac{|\mathcal{W}(n,k,A)|}{m^n} = 1$.*

Proof: Let $L_{\alpha_1\alpha_2\ldots\alpha_k}$ be the set of words of length n over A not containing the word $\alpha_1\alpha_2\ldots\alpha_k$ as a subword. It is clear that

$$\mid \mathcal{W}(n,k,A) \mid = m^n - \mid \bigcup_{(\alpha_1,\ldots,\alpha_k)} L_{\alpha_1\alpha_2\ldots\alpha_k} \mid$$

and by Lemma 1.1,

$$\mid \bigcup_{(\alpha_1,\ldots,\alpha_k)} L_{\alpha_1\alpha_2\ldots\alpha_k} \mid \leq \sum_{(\alpha_1,\ldots,\alpha_k)} \mid L_{\alpha_1\alpha_2\ldots\alpha_k} \mid \leq m^k L(n) \tag{7}$$

By (7) it is sufficient to prove that $lim_{n\to\infty}L(n)/m^{n-k} = 0$, provided $lim_{n\to\infty}k(n)/logn = 0$. By Lemma 2.1 if $n \geq 3$ we have $| x_i | \leq 1$, hence $| (k+1)x_i - km | \geq | (k+1) | x_i | -km | = km - (k+1) | x_i | > km$ and $| x_i(x_i - 1) | \leq | x_i - 1 | \leq 2$. It follows that $| C_i | \leq \frac{2}{km(m-1)}$ for every $i \geq 3$.
Since $x_2 > m - \frac{1}{m^{k-1}}$ we deduce that $(k+1)x_2 - km > m - \frac{k+1}{m^{k-1}} \geq \frac{m}{4}$ for every $k, m \geq 2$. It follows that $C_2 \leq \frac{4(x_2^2 - x_2)}{m(m-1)} < \frac{4m}{m-1}$ because $x_2 < m$. Hence $| C_i | \leq 8$ for every $2 \leq i \leq k+1$ and $m \geq 2$. By Lemma 1.3 we get $\frac{|L(n)|}{m^{n-k}} \leq \frac{8kx_2^n}{m^{n-k}}$ and $lim_{n\to\infty}km^k(\frac{x_2}{m})^n = e^{lim_{n\to\infty}((\frac{x_2}{m}-1)n+klnm+lnk)}$ since iv) of Lemma 1.2 holds. By denoting $f(n) = (\frac{x_2}{m} - 1)n + klnm + lnk$, by iv) we deduce $f(n) < -\frac{n}{m^{k+1}} + klnm + lnk < -\frac{n}{m^{k+1}} + 2klnm$. But $log\frac{n}{m^{k+1}} = logn(1 - \frac{k+1}{logn}logm) \to \infty$ as $n \to \infty$ because $\frac{k}{logn} \to 0$; also $log\frac{km^{k+1}}{n} = logk + (k+1)logm - logn \to -\infty$ as $n \to \infty$. Consequently, $lim_{n\to\infty}f(n) = -\infty$, which implies $lim_{n\to\infty}L(n)/m^{n-k} = e^{-\infty} = 0$. □

Note that for fixed k the property expressed in Theorem 2.1 is well-known and it follows e.g. from the property that almost all random strings (in both Kolmogorov-Chaitin and Chaitin senses) satisfy various conditions of normality (first introduced by Borel) (see [1], [2]).

Acknowledgement

I would like to thank Cristian Calude who suggested the study of this problem for stimulating discussions.

References

[1] C. Calude, *Information and randomness: An algorithmic perspective*, Springer-Verlag, New York Heidelberg Berlin, 1995.

[2] C. Calude, I. Chiţescu, Qualitative properties of P. Martin-Löf random sequences, *Bolletino della Unione Matematica Italiana* (7), 3-B(1989), 229-240.

[3] L. Priese, personal e-mail communication, May 2, 1995.

[4] I. Proskuriakov, *Recueil de problèmes d'algèbre linéaire*, Mir, Moscou, 1989.