Canad. Math. Bull.Vol. **35** (2), 1992 pp. 161–166

# WORDS WITHOUT NEAR-REPETITIONS

J. CURRIE AND A. BENDOR-SAMUEL

ABSTRACT. We find an infinite word $w$ on four symbols with the following property: Two occurrences of any block in $w$ must be separated by more than the length of the block. That is, in any subword of $w$ of the form $xyx$, the length of $y$ is greater than the length of $x$. This answers a question of C. Edmunds connected to the Burnside problem for groups.

1. **Introduction.** In their solution of the Burnside problem for groups [5], Novikov and Adjan use a result from combinatorics on words:

> There is an infinite word $v$ on the alphabet $\{0, 1\}$ such that $v$ contains no subword of the form $xxx$, $x \neq \epsilon$. [2,6]

Novikov and Adjan invoke this result at the end of their notoriously long and involved proof. The bulk of their proof, filling a book of 300+ pages, involves constructions of groups. C. Edmunds [4] suggests that it may be possible to find a shorter proof by using stronger results from combinatorics on words, rather than by finding new group theoretic constructions. With this motivation, Edmunds poses the following question:

Can one find a finite alphabet $S$, and some infinite word $w$ over $S$ such that whenever $xyx$ is a subword of $w$, the length of $y$ is greater than the length of $x$?

We answer Edmunds' question in the affirmative. The smallest alphabet for which such a $w$ can exist is a 4 letter alphabet.

2. **Notation.** Our notation follows the usual notation of automata theory. Let $S$ be a set. A *word* is a finite sequence of elements of $S$. We refer to $S$ as an *alphabet*, its elements as *letters*. The set of all words over $S$ is denoted $S^*$. We take a naive view of words as strings of letters; thus the concatenation of two words $w$ and $v$, written $wv$, is simply the string of letters consisting of the letters of $w$ followed by the letters of $v$.

Say that $v$ is a *subword* of $w$ if we can write $w = uvz$; $u, v, z \in S^*$. If $w = uv$ then we say that $u$ is a *prefix* of $w$; $v$ is a *suffix* of $w$. The *empty word*, denoted $\epsilon$, is the word with no letters in it. Denote by $|w|$ the *length* of $w$, equal to the number of letters of $w$.

Let $S, T$ be alphabets. A *substitution* $h: S^* \longrightarrow T^*$ is a function generated by its values on $S$. That is, suppose $w \in S^*$, $w = a_1a_2\cdots a_m$; $a_i \in S$ for $i = 1$ to $m$. Then $h(w) = h(a_1)h(a_2)\cdots h(a_m)$.

---

The research of the first author was supported by an NSERC Operating Grant.
The second author was supported by an NSERC Undergraduate Summer Research Award.
Received by the editors February 13, 1991.
AMS subject classification: 68Q, 03C.

Let $S$ be an alphabet, $w \in S^*$ a word over $S$. If we can write $w = uxyxv$ with $|y| \leq |x|$, $u, v, x, y \in S^*$, we call $w$ *near-repetitive*, and call $xyx$ a *near-repetition*. If $w$ is not near-repetitive, call $w$ *varied*.

3. **Construction of varied words.** By König's Infinity Lemma, to show that there is an infinite varied word over a finite alphabet $S$, it suffices to show that there are arbitrarily long varied words over $S$. Let $S$ be the alphabet $S = \{1, 2, 3, 4, 5\}$. Consider the substitution $f: S^* \longrightarrow S^*$ given by

$$f(1) = 123145213412435$$
$$f(2) = 123154234531425$$
$$f(3) = 123152413425324$$
$$f(4) = 123143254135245$$
$$f(5) = 123153452132534.$$

We will prove that $f^n(1)$ is varied. To begin, we make some observations concerning $f$:

OBSERVATION 1. We see that $f$ replaces each letter of $S$ by a string of fifteen letters. Thus if $u \in S^*$, $|f(u)| = 15|u|$. ■

OBSERVATION 2. The images of different letters under $f$ can have a common suffix of length at most 1. That is, suppose that $u, v \in S$ and we have

$$f(u) = UW, f(v) = VW, \ |W| \geq 2.$$

Then $u = v$. ■

One concludes from Observation 2 that $f$ is 1-1.

OBSERVATION 3. The images of different letters under $f$ can have a common prefix of length at most 5. Thus suppose that $u, v \in S$ and we have

$$f(u) = WU'', f(v) = WV'', \ |W| \geq 6.$$

It follows that $u = v$. ■

OBSERVATION 4. The images of different letters under $f$ can have a common subword of length at most 6. In fact, suppose that $u, v \in S$ and we have

$$f(u) = U'WU'', f(v) = V'WV'', \ |W| \geq 7.$$

We must have $U' = V'$, $U'' = V''$, $u = v$. ■

OBSERVATION 5. Call a word $w$ a *suffix-prefix* if we can write $w = uv$ where $u$ is the non-empty suffix of the image of some letter under $f$, and $v$ is the non-empty prefix of the image of some letter. Note that no non-empty prefix of the image of a letter is the suffix of the image of a letter. Thus if $w$ can be expressed as a suffix-prefix then the words $u$ and $v$ are unique.

https://doi.org/10.4153/CMB-1992-023-6 Published online by Cambridge University Press

The longest instance of a suffix-prefix in the image under $f$ of a letter is 3412 in $f(1)$. Thus if $u, v, w \in S$ and

$$f(u) = U'V''W'U'',\ f(v) = V'V'',\ f(w) = W'W'',\ \text{with } W', V'' \neq \epsilon,$$

then $|V''W'| \leq 4$. ∎

Using some of these observations we prove the following lemma.

LEMMA. *Let $u = u_1u_2\cdots u_m$, $v = v_1v_2\cdots v_n$ with the $u_i, v_j \in S$. Let $f(u_i) = U_i$, $f(v_i) = V_i$. Suppose that for some word $w$ we can write*

$$f(u) = U_1U_2\cdots U_j'wU_k''U_{k+1}\cdots U_m$$

*and*

$$f(v) = V_1V_2\cdots V_s'wV_t''V_{t+1}\cdots V_n, \quad |w| \geq 7$$

*where*

$$U_j = U_j'U_j'',\ U_k = U_k'U_k'',\ V_s = V_s'V_s'',\ V_t = V_t'V_t''.$$

*Then*

$$|U_j'| \equiv |V_s'| \pmod{15},\ |U_k''| \equiv |V_t''| \pmod{15}.$$

PROOF. By Observation 1, it follows that

$$|U_j'| + |w| + |U_k''| \equiv |V_s'| + |w| + |V_t''| \equiv 0 \pmod{15}.$$

It thus suffices to show that $U_j' \equiv V_s' \pmod{15}$. To do this, we will assume that $|w| = 7$, replacing $w$ by its first 7 letters if necessary. It follows that $k \leq j+1$, $t \leq s+1$. We will also assume without loss of generality that $|U_j'|, |V_s'|, |U_k''|, |V_t''| < 15$. The word $w$ is thus a subword of $U = U_jU_{j+1}$ and of $V = V_sV_{s+1}$.

Suppose that $w$ is not a suffix-prefix. Then $w$ must be a subword of either $U_j$ or $U_{j+1}$. Assume first that $w$ is a subword of $U_j$. Again, $w$ must be a subword of either $V_s$ or $V_{s+1}$. If $w$ is a subword of $V_s$, then Observation 4 implies that $|U_j'| = |V_s'|$, and we are done. Otherwise, $w$ is a prefix of $V_{s+1}$, and $|V_s'| = 0$. By Observation 4, $w$ is also a prefix of $U_j$, so that $U_j' = \epsilon = V_s'$. (In this case $j = k$.) A symmetrical argument deals with the possibility that $w$ is a subword of $U_{j+1}$.

Suppose then that $w$ is a suffix-prefix, $w = U_j''U_{j+1}' = V_s''V_{s+1}'$. It follows from Observation 5 that $U_j' = V_s'$. ∎

THEOREM 1. *For all $n \in \mathbb{N}$, the word $f^n(1)$ is varied.*

PROOF. We proceed by induction. One checks that $f^1(1) = f(1)$ is varied. Let $n$ be least such that $f^n(1)$ is near repetitive. Let $e = e_1e_2\cdots e_m$ be a subword of $f^{n-1}(1)$ of minimal length such that $f(e)$ contains a near repetition $xyx$, $|y| \leq |x|$. It is convenient to make two cases:

https://doi.org/10.4153/CMB-1992-023-6 Published online by Cambridge University Press

CASE 1. We have $|x| \leq 6$.

In this case, $|xyx| \leq 18$. It follows that $|e| \leq 3$. Moreover, $e$ is a varied word since it is a subword of $f^{n-1}(1)$. To show the impossibility of this case, it suffices to check that $f(e)$ is varied whenever $e \in S^*$ is varied and $|e| = 3$. Such a word $e$ must consist of three distinct letters, and one checks that the relevant 60 words are varied.

CASE 2. We have $|x| \geq 7$. We may also assume, by our disposition of case 1, that $m \geq 4$.

Let $f(e_i) = E_i$ and write $f(e) = E_1'xyxE_m'' = E_1'xE_j''E_{j+1}\cdots E_m = E_1\cdots E_k'xE_m''$, where $E_1 = E_1'E_1''$, $E_j = E_j'E_j''$, $E_k = E_k'E_k''$, $E_m = E_m'E_m''$ and $E_1''$, $E_j'$, $E_k''$, $E_m'$ are non-empty. (We know that $E_1''$ and $E_m'$ are non-empty by the minimality of $|e|$. Let the others be non-empty by a notational convention.) We must have $j < m$. Otherwise $E_2E_3$ is a subword of our first occurrence of $x$, but the second occurence of $x$ is a subword of $E_m$. This is a contradiction on the length of $x$. Also, $k < m$. Otherwise the second occurrence of $x$ is a subword of $E_m$, but $E_1''E_2E_3$ is a subword of $xy$. This gives the contradiction $30 < |E_1''E_2E_3| \leq |xy| \leq 2|x| \leq 2|E_m| = 30$. Similarly, $1 < j \leq k < m$.

By the lemma, $|E_1'| \equiv |E_k'|$, $|E_j''| \equiv |E_m''|$ (mod 15). Since $E_1''$, $E_j'$, $E_k''$, $E_m'$ are non-empty, the congruence can in fact be replaced by equality. Without loss of generality, we may assume that $|E_1''| \leq 1$. Suppose not. Then $|E_1''| = |E_k''| \geq 2$. Since $E_1''$ and $E_k''$ are prefixes of $x$, and have the same length they are equal. It follows from Observation 3 that $e_1 = e_k$.

Write $x = x'x''$ where $|x''| = \max(0, |E_1'| - |y|)$. If $|y| > |E_1'|$, then write $y = \hat{y}y''$ where $|y''| = |E_1'|$. Otherwise, let $\hat{y} = \epsilon$. We see that $f(e)$ contains the near repetition $\hat{x}\hat{y}\hat{x}$, where $\hat{x} = E_1'x'$. If we replace $x$ by $\hat{x}$, and $y$ by $\hat{y}$ in our argument, we get $|E_1| = 0$. (In other words, we extend both the occurrences of our original $x$ by adding a prefix $E_1' = E_k'$ in front. In the case of the second $x$, this will shorten $y$ by $|E_1'|$. If $|y|$ is shorter than $|E_1'|$, an amount $|E_1'| - |y|$ is removed from the end of each $x$, and $y$ disappears.) Similarly, without loss of generality, we may assume that $|E_m'| \leq 5$.

We can write

$$x = E_1''E_2\cdots E_j' = E_k''E_{k+1}\cdots E_m'.$$

In fact, $E_1'' = E_k''$, $E_j'' = E_m''$, $E_2E_3\cdots E_{j-1} = E_{k+1}E_{k+2}\cdots E_{m-1}$. Since $f$ is 1-1, we have $e_2\cdots e_{j-1} = e_{k+1}\cdots e_{m-1}$.

Let $a = e_2\cdots e_{j-1} = e_{k+1}\cdots e_{m-1}$, $b = e_j\cdots e_k$. We claim that $aba$ is a near repetition in $e$; that is, that $|b| \leq |a|$. This will be a contradiction, for $e$ must be varied. If $j = k$ the claim is clearly true. Otherwise,

$$\begin{aligned}
|a| = |e_2\cdots e_{j-1}| &= (|E_2\cdots E_{j-1}|)/15 \\
&= \big(|x| - (|E_1''| + |E_j'|)\big)/15 \\
&\geq \big(|x| - (1+5)\big)/15 \\
&= (|x| - 6)/15,
\end{aligned}$$

https://doi.org/10.4153/CMB-1992-023-6 Published online by Cambridge University Press

$$\begin{aligned}|b| = |e_j \cdots e_k| &= (|E_j \cdots E_k|)/15 \\ &= \big(|y| + (|E_j'| + |E_k''|)\big)/15 \\ &\leq (|x| + 6)/15.\end{aligned}$$

It follows that $|b| - |a| \leq 12/15$. Since $|a|$ and $|b|$ are integers, we conclude that $|b| \leq |a|$. ∎

One discovers quickly that the longest varied words over the alphabet $\{1, 2, 3\}$ are permutations of 1231. Thus there is no infinite varied word on a 3 letter alphabet. Let $T = \{1, 2, 3, 4\}$, and let $g: T^* \longrightarrow T^*$ be given by

$$\begin{aligned}g(1) =& 123421432413423124321341231421324123421431241321423124\\ & 321341231432413421431234132142312413421432412314213243\\ g(2) =& 123421432413423124321423413243123421324123142134124231\\ & 423124132143123413243142134123143213423124321423413243\\ g(3) =& 123421432413423143213412314213243123413214312413421432\\ & 412314213412431423413243123421324123143213412431421324\\ g(4) =& 123421432413423143213412431423413214312413421432412342\\ & 132431423412432134231432413421431241321423412431421324\end{aligned}$$

THEOREM 2. *The word $g^n(1)$ is varied for every $n \in \mathbb{N}$.*

This theorem is proved analogously to Theorem 1, with proportionately more checking. We see that $g$ replaces each letter of $T$ by a string of 108 letters. The images of different letters under $g$ can have a common suffix of length at most 13, a common prefix of length at most 24. With similar observations and proceeding as in the previous theorem, one establishes a lemma:

LEMMA. *Let $u = u_1u_2 \cdots u_m$, $v = v_1v_2 \cdots v_n$ with the $u_i, v_j \in S$. Let $g(u_i) = U_i$, $g(v_i) = V_i$. Suppose that for some word $w$ we can write*

$$g(u) = U_1U_2 \cdots U_j'wU_k''U_{k+1} \cdots U_m \text{ and } g(v) = V_1V_2 \cdots V_s'wV_t''V_{t+1} \cdots V_n,$$

$|w| \geq 38$ *where*

$$U_j = U_j'U_j'',\ U_k = U_k'U_k'',\ V_s = V_s'V_s'',\ V_t = V_t'V_t''$$

*Then*

$$|U_j'| \equiv |V_s'| \pmod{108},\ |U_k''| \equiv |V_t''| \pmod{108}.$$

∎

The proof of Theorem 2 is similar to that of Theorem 1. In the final phase, the proof of Theorem 1 depended on an inequality involving the quantities in Observations 1, 2 and 3: $1 + 5 < 15/2$. In Theorem 2, we have the analogous inequality: $13 + 24 < 108/2$.

We have thus answered Edmunds' question in the affirmative, and shown that a four letter alphabet is the smallest on which infinite varied words exist.

## REFERENCES

**1.** S. I. Adian, *The Burnside Problem and Identities in Groups*, Springer-Verlag, New York 1979.

**2.** S. Arshon, *Demonstration de l'existence des suites asymetriques infinies*, Mat. Sb. **2** (44) (1937), 769–779.

**3.** T. C. Brown, *Is there a sequence on four symbols in which no two adjacent segments are permutations of each other?*, Amer. Math. Monthly **78**(1971), 886–888.

**4.** C. Edmunds, personal communication.

**5.** P. S. Novikov and S. I. Adjan, *Infinite periodic groups I, II, III*, Math. U.S.S.R. Izv. **2**(1968) I: 209–236, II: 241–479, III: 665–685.

**6.** A. Thue, *Uber unendliche Zeichenreihen*, Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana (1912), 1–67.

*Department of Mathematics*
*University of Winnipeg*
*Winnipeg, Manitoba*
*R3B 2E9*