BULL. AUSTRAL. MATH. SOC. 94A10
VOL. 14 (1976), 249-258.

The comma-free codes with words of length two

A.H. Ball and L.J. Cummings

A code not requiring a distinct symbol to separate words is called comma-free. Two codes are isomorphic if one can be obtained from the other by a permutation of the underlying alphabet. Since subcodes of comma-free codes are comma-free, we investigate only maximal comma-free codes. All isomorphism classes of maximal comma-free codes with words of length 2 are determined and a natural representative of each class is given.

1. Introduction

Let $\Sigma_n = \{0, 1, \ldots, n-1\}$ be an alphabet of n symbols. Let F be a block code of word length k over Σ_n. The code F is said to be comma-free if whenever

$$a_1 \ldots a_k \in F \text{ and } b_1 \ldots b_k \in F$$

then the words

$$a_2 \ldots a_k b_1, a_3 \ldots a_k b_1 b_2, \ldots, a_k b_1 \ldots b_{k-1} \tag{1}$$

are *not* in F. The words (1) are called the *overlaps* of $a_1 \ldots a_k$ and $b_1 \ldots b_k$. Alternately, a block code is comma-free if a distinct symbol is not required to separate code words in a message. Subcodes of comma-free codes are comma-free.

Mathematical study of comma-free codes was initiated in 1958 by

Received 2 December 1975.

https://doi.org/10.1017/S0004972700025065 Published online by Cambridge University Press

Golomb, Gordon and Welch [2]. These authors gave an upper bound for the size $W_k(n)$ of a maximal comma-free code as a function of the alphabet length n and the block length k. They conjectured that the bound was always attained when the k was an *odd* integer. Eastman [1] found a construction resolving their conjecture 6 years later.

When the block length is even, less is known about $W_k(n)$. Golomb, Gordon and Welch [2] proved that for all positive integers n,

$$W_2(n) = \left[\frac{n^2}{3}\right] \tag{2}$$

where $[\cdot]$ denotes the greatest integer function. Maximal comma-free codes over a binary alphabet have been constructed for $k = 2, 4, 6, 8$, and 10 in [2], [3], and [4]. In this paper we determine all maximal comma-free codes with words of length 2 over arbitrary finite alphabets.

2. Preliminaries

Throughout the remainder of this paper all codes will be assumed to have words of length 2. Every comma-free code F over an alphabet Σ induces a partition of Σ as follows:

$$A(F) = \{a \in \Sigma \mid a \text{ only begins words of } F\},$$

$$B(F) = \{b \in \Sigma \mid b \text{ both begins and ends words of } F\},$$

$$C(F) = \{c \in \Sigma \mid c \text{ only ends words of } F\}.$$

We say that F *has parameters* (q_1, q_2, q_3) if

$$|A(F)| = q_1, \quad |B(F)| = q_2, \quad |C(F)| = q_3,$$

where $|X|$ always denotes the cardinality of a set X. Maximality of a comma-free code restricts the parameters considerably as the following lemma shows. The authors are indebted to Dr T.K. Sheng for the proof of this lemma.

LEMMA 1. *Let n be a positive integer. The only common positive integral solutions of the equations*

$$xy + yz + zx = \left[\frac{n^2}{3}\right] \tag{3}$$

https://doi.org/10.1017/S0004972700025065 Published online by Cambridge University Press

and

(4) $$x + y + z = n$$

are

$$(q, q, q) \quad n = 3q ,$$

$$(q+1, q, q) \quad n = 3q + 1 ,$$

$$(q+1, q+1, q) \quad n = 3q + 2 ,$$

and their permutations.

Proof. Squaring (4) we obtain

(5) $$xy + yz + xz = \frac{n^2}{3} - \frac{1}{6}\left[(x-y)^2+(y-z)^2+(z-x)^2\right]$$

We have

(6) $$\left[\frac{n^2}{3}\right] = \begin{cases} 3q^2 & n = 3q , \\ 3q^2 + 2q & n = 3q + 1 , \\ 3q^2 + 4q + 1 & n = 3q + 2 . \end{cases}$$

Combining (5) and (6) with (3) we obtain

(7) $$(x-y)^2 + (y-z)^2 + (z-x)^2 = \begin{cases} 0 & n = 3q , \\ 2 & \text{otherwise.} \end{cases}$$

If $n = 3q$ then (7) implies $x = y = z = q$. Otherwise, exactly two of the differences $|x-y|$, $|y-z|$, and $|z-x|$ are 1. Without loss of generality assume $x - y = 1$ and further suppose $n = 3q + 1$. If either $y - z = 1$ or $z - y = 1$ we obtain contradictions of (7) and (4) respectively. Similarly, if $z - x = 1$ we contradict (7). Therefore $x - z = 1$ so that (4) yields $z = q$, implying $x = q + 1$, $y = q$, and $z = q$. The argument is similar in case $n = 3q + 2$.

If π is a permutation of the alphabet Σ_n then π induces the natural mapping $f(\pi) : \Sigma_n \times \Sigma_n \to \Sigma_n \times \Sigma_n$ defined by

$$f(\pi)(xy) = \pi(x)\pi(y) ,$$

where the ordered pairs (x, y) of the Cartesian product $\Sigma_n \times \Sigma_n$ are written as simply xy for notational convenience. For any permutation π, the image of a comma-free code under $f(\pi)$ is comma-free. Two codes F and F' over the same alphabet Σ_n are *isomorphic* if there is a permutation π of Σ_n such that $f(\pi)$ is a one-to-one mapping of F onto F'. Isomorphic codes have the same parameters.

The mapping $xy \mapsto yx$ also preserves the comma-free property of codes with words of length 2. The image of a code under this mapping is called its *transpose*. If F has parameters (q_1, q_2, q_3) then the transpose of F has parameters (q_3, q_2, q_1). An obvious necessary condition that a code with parameters (q_1, q_2, q_3) be isomorphic to its transpose is $q_1 = q_3$. For maximal comma-free codes this condition is also sufficient and in Corollary 6 we determine precisely which of these codes are isomorphic to their transposes.

LEMMA 2. *Let F be a comma-free code over Σ_n with parameters (q_1, q_2, q_3) where $q_i > 1$, $i = 1, 2, 3$. Choose*

(8) $$a \in A(F) \,, \quad b \in B(F) \,, \quad c \in C(F) \,.$$

If all words of F containing at least one of a, b, or c are deleted then the remaining words form a comma-free code F_1 and

(9) $$|F_1| \geq |F| - 2n + 3 \,.$$

Proof. There are at most $q_1 + q_2$ words of the form xc in F since x can be chosen only from $A(F)$ or $B(F)$. There are at most $q_2 + q_3$ words of the form ay in F since y can be chosen only from $B(F)$ or $C(F)$. There are at most q_3 words of the form by since words b_1b_2 with $b_1, b_2 \in B(F)$ cannot appear in comma-free codes. Similarly there are at most q_1 words of the form xb in F. Thus, at most

$$(q_1+q_2) + (q_2+q_3-1) + (q_1-1) + (q_3-1) = 2(q_1+q_2+q_3) - 3 = 2n - 3$$

words of F are deleted. Therefore, $|F_1| \geq |F| - 2n + 3$.

https://doi.org/10.1017/S0004972700025065 Published online by Cambridge University Press

3. Results

For notational convenience we write for integers x, y, a, b ,

$$xy \equiv ab \pmod 3$$

if $x \equiv a \pmod 3$ and $y \equiv b \pmod 3$

THEOREM 3. *If* $n \equiv 0 \pmod 3$ *then every maximal comma-free code with words of length* 2 *over* Σ_n *is isomorphic to*

$$C_n = \{xy \mid xy \equiv 01, 02, \text{ or } 12 \pmod 3 \text{ where } x, y \in \Sigma_n\}$$

or C_n *transposed.*

Proof. Let F be a maximal comma-free code over Σ_{3q} . The proof is by induction on q .

If $q = 1$ then F has size 3 by (2). The maximality of F implies every element of Σ_3 appears in some word of F . If 0 does not begin a word of F then consider the transpose of F instead. We may suppose F contains a word xy with $x = 0$. Since F is comma-free, y cannot be 0 ; so y is either 1 or 2 . Therefore F contains either 01 or 02 .

If F contains both 01 and 02 then F contains either 12 or 21 . This follows because then F cannot contain either 10 or 20 and remain comma-free, leaving only 12 or 21 as possibilities. If F contains 21 then the permutation (12) of Σ_3 induces an isomorphism of F and C_3 . Otherwise, F is identically C_3 .

Now suppose F contains only 01 . The remaining words of F are among 20, 12 , and 21 . Since the pair 12 and 21 cannot appear in a comma-free code, F must contain 20 . But {01, 20, 12} is not comma-free because 12 is an overlap of 01 and 20 . Therefore, $F = \{01, 20, 21\}$. The permutation (012) of Σ_3 induces an isomorphism of F and C_3 .

The argument is similar if F contains only 02 .

Now suppose $q > 1$. By Lemma 1, F has parameters (q, q, q) . Choosing a, b, c as in (8) and deleting words of F containing them, we obtain a comma-free subcode F_1 of F. F_1 has parameters $(q-1, q-1, q-1)$. Since $|F| = 3q^2$, (9) implies $|F_1| \geq 3(q-1)^2$. But the number of words in any maximal comma-free code over Σ_{3q-3} is $3(q-1)^2$ by (2). Therefore F_1 is a maximal comma-free code. By induction, there is an isomorphism $f(\pi_1)$ of F_1 and C_{3q-3} or C_{3q-3} transposed. The permutation π_1 of Σ_{3q-3} may be extended to a permutation π of Σ_{3q} by defining

$$\pi(a) = 3q - 3 , \quad \pi(b) = 3q - 2 , \quad \pi(c) = 3q - 1 .$$

We verify that $f(\pi)$ is an isomorphism of F and C_{3q} as follows.

For words containing a, b or c we have

$$f(\pi)(ay) = (3q-3)\pi(y) , \quad f(\pi)(xb) = \pi(x)(3q-2) ,$$

$$f(\pi)(by) = (3q-2)\pi(y) , \quad f(\pi)(xc) = \pi(x)(3q-1) .$$

If, for example,

$$\pi(x)(3q-1) = \pi(x')(3q-1)$$

obviously $x = x'$. If, say,

$$\pi(x)(3q-1) = (3q-3)\pi(y) ,$$

then $x = a$ and $y = c$. The other cases are similar. Therefore $f(\pi)$ is one-to-one.

THEOREM 4. *If* $n \equiv 1 \pmod 3$ *then every maximal comma-free code with words of length* 2 *over* Σ_n *is isomorphic to* C_n,

$$D_n = \{xy \mid xy \equiv 10, 02, \text{ or } 12 \pmod 3 \text{ where } x, y \in \Sigma_n\} ,$$

or one of their transposes.

Proof. Let F be a maximal comma-free code over Σ_{3q+1}. The proof is by induction on q.

If $q = 1$ then F has size 5 by (2). By Lemma 1 we may suppose F has parameters $(2, 1, 1)$, $(1, 2, 1)$ or $(1, 1, 2)$. If F has parameters $(1, 1, 2)$ then the transpose of F would have parameters $(2, 1, 1)$.

https://doi.org/10.1017/S0004972700025065 Published online by Cambridge University Press

If F has parameters $(2, 1, 1)$ and $a \in A(F)$ then there are at most 2 words in F containing a. When the words containing a are deleted from F the resulting comma-free code F_1 has at least 3 words. Therefore (2) implies that F_1 contains exactly 3 words and so is a maximal comma-free code. By Theorem 3 there is an isomorphism $f(\pi_1)$ of F_1 and C_3 or C_3 transposed. The permutation π_1 of Σ_3 can be extended to a permutation π of Σ_4 by setting $\pi(a) = 3$. For the two words of the form ax in F we obtain

$$f(\pi)(ax) = 3\pi(x)$$

so that $f(\pi)$ is one-to-one.

If F has parameters $(1, 2, 1)$ and $b \in B(F)$ then deletion of words containing b leads again to an isomorphism of the remaining subcode with C_3 or its tranpose. The permutation (01) induces an isomorphism of C_3 and the subcode $\{10, 02, 12\}$ of D_4. The permutation (02) induces an isomorphism of C_3 and its transpose. Extending the appropriate permutation to a permutation π of Σ_4 by assigning $\pi(b) = 3$, we obtain an isomorphism of F and D_4.

Now suppose $q > 1$. By Lemma 1, F has parameters $(q+1, q, q)$, $(q, q+1, q)$, or $(q, q, q+1)$. If F has parameters $(q, q, q+1)$ then F transposed has parameters $(q+1, q, q)$.

Assume F has parameters $(q+1, q, q)$. Choosing a, b, and c as in (8) and deleting the words containing them, we obtain a comma-free subcode F_1 of F. The code F_1 has parameters $(q, q-1, q-1)$. Since F has $3q^2 + 2q$ words by (2), the inequality (9) implies $|F_1| \geq 3q^2 - 4q + 1$. But the size of a maximal comma-free code over Σ_{3q-2} is $3q^2 - 4q + 1$ by (2) and (6). Therefore F_1 is a maximal comma-free code over Σ_{3q-2}

By induction, there is a permutation π_1 of Σ_{3q-2} inducing an

https://doi.org/10.1017/S0004972700025065 Published online by Cambridge University Press

isomorphism of F_1 with C_{3q-2}, D_{3q-2}, or one of their transposes. But D_{3q-2} and its transpose have parameters $(q-1, q, q-1)$ and F_1 has parameters $(q, q-1, q-1)$. Therefore F_1 is isomorphic with C_{3q-2}. We extend π_1 to a permutation π of Σ_{3q+1} by defining

$$\pi(a) = 3q\ , \quad \pi(b) = 3q - 2\ , \quad \pi(c) = 3q - 1\ .$$

It is readily verified that $f(\pi)$ is an isomorphism of F and C_{3q+1}.

Now assume F has parameters $(q, q+1, q)$. As above we obtain a subcode F_1 of F which is maximal and comma-free over Σ_{3q-2} but with parameters $(q-1, q, q-1)$ in this case. This time induction implies F_1 is isomorphic to D_{3q-2} or its transpose via a permutation π_1 of Σ_{3q-2}. The assignments

$$\pi(a) = 3q - 1\ , \quad \pi(b) = 3q\ , \quad \pi(c) = 3q - 2\ ,$$

extend π_1 to a permutation π of Σ_{3q+1} which induces an isomorphism of F and D_{3q+1}.

THEOREM 5. *If* $n \equiv 2 \pmod 3$ *then every maximal comma-free code with words of length* 2 *over* Σ_n *is isomorphic to* C_n,

$$E_n = \{xy \mid xy \equiv 01,\ 02,\ \text{or}\ \ 21 \pmod 3)\ \text{where}\ \ x, y \in \Sigma_n\}\ ,$$

or one of their transposes.

Proof. Let F be a maximal comma-free code over Σ_{3q+2}. The proof is by induction on q.

If $q = 1$ then F has size 8 by (2). By Lemma 1, we may suppose F has parameters $(2, 2, 1)$, $(2, 1, 2)$, or $(1, 2, 2)$. If F has parameters $(1, 2, 2)$ then the transpose of F would have parameters $(2, 2, 1)$.

If F has parameters $(2, 2, 1)$ and $b \in B(F)$ we delete the words of F containing b to obtain a subcode F_1 with parameters $(2, 1, 1)$. There can be at most 2 words of the form xb in F and only one of the form by. Therefore, F_1 contains at least 5 words. But F_1 contains

https://doi.org/10.1017/S0004972700025065 Published online by Cambridge University Press

at most 5 words by (2). By Theorem 4, F_1 is isomorphic to C_4 since C_4 transposed has parameters (1, 1, 2) and both D_4 and its transpose have parameters (1, 2, 1).

The permutation inducing this isomorphism may be extended to a permutation π of Σ_5 by the assignment $\pi(b) = 4$. This induces an isomorphism of F and C_5.

On the other hand, if F has parameters (2, 1, 2) and $c \in C(F)$ then deletion of words containing c leaves a subcode F_1 also having parameters (2, 1, 1) and isomorphic to C_4. This time, however, F cannot be isomorphic to C_5 because the assignment $\pi(c) = 4$ could not induce correspondences with the words 04, 42, and 34 in C_5. Since C_4 and E_4 are isomorphic via the permutation $\pi = (12)$, defining $\pi(c) = 4$ will induce an isomorphism of F and E_5.

Now suppose $q > 1$. By Lemma 1, F has parameters $(q+1, q+1, q)$, $(q+1, q, q+1)$ or $(q, q+1, q+1)$. If F has parameters $(q, q+1, q+1)$ then F transposed has parameters $(q+1, q+1, q)$.

Assume F has parameters $(q+1, q+1, q)$. Choosing a, b, and c as in (8) and deleting the words containing them, we obtain a comma-free subcode F_1 of F. The code F_1 has parameters $(q, q, q-1)$. By (9), $|F_1| \geq 3q^2 - 2q$. But the size of a maximal comma-free code over Σ_{3q-1} is $3q^2 - 2q$ by (2). Therefore F_1 is isomorphic to C_{3q-1}, E_{3q-1}, or one of their transposes.

But E_{3q-1} and its transpose have parameters $(q, q-1, q)$ while C_{3q-1} has parameters $(q, q, q-1)$. Therefore F_1 is isomorphic to C_{3q-1}, via a permutation π_1 of Σ_{3q-1}. Extending π_1 to a permutation π of Σ_{3q+2} by defining

$$\pi(a) = 3q, \quad \pi(b) = 3q + 1, \quad \pi(c) = 3q - 1,$$

we obtain an isomorphism of F and C_{3q+2}.

https://doi.org/10.1017/S0004972700025065 Published online by Cambridge University Press

On the other hand, if F has parameters $(q+1, q, q+1)$ similar arguments lead to a subcode F_1 isomorphic to either E_{3q-1} or its transpose. Extending the permutation π_1 of Σ_{3q-1} involved to Σ_{3q+2} by defining

$$\pi(a) = 3q \ , \quad \pi(b) = 3q - 1 \ , \quad \pi(c) = 3q + 1 \ ,$$

we obtain an isomorphism of F with either E_{3q+2} or its transpose.

COROLLARY 6. *The maximal comma-free codes with words of length* 2 *which are isomorphic to their transposes are* C_{3q}, D_{3q+1}, *and* E_{3q+2} *for each positive integer* q .

Proof. The necessary condition that $q_1 = q_3$ if the code has parameters (q_1, q_2, q_3) is also sufficient. This follows because each of the q_1 elements of the set A appear with equal frequency in the code and similarly for the set C .

References

[1] Willard L. Eastman, "On the construction of comma-free codes", *IEEE Trans. Information Theory* IT-11 (1965), 263-267.

[2] S.W. Golomb, Basil Gordon and L.R. Welch, "Comma-free codes", *Canad. J. Math.* 10 (1958), 202-209.

[3] B.H. Jiggs, "Recent results in comma-free codes", *Canad. J. Math.* 15 (1963), 178-187.

[4] Yoji Niho, "On maximal comma-free codes", *IEEE Trans. Information Theory* IT-19 (1973), 580-581.

Department of Mathematics,
University of Newcastle,
Newcastle, New South Wales;

Faculty of Mathematics,
University of Waterloo,
Waterloo, Ontario, Canada.