RECENT RESULTS IN COMMA-FREE CODES

B. H. JIGGS

1. Introduction. A set D of k-letter words is called a *comma-free dictionary* **(2)**, if whenever $(a_1a_2 \ldots a_k)$ and $(b_1b_2 \ldots b_k)$ are in D, the "overlaps" $(a_2a_3 \ldots a_kb_1)$, $(a_3a_4 \ldots a_kb_1b_2), \ldots, (a_kb_1 \ldots b_{k-1})$ are not in D. We say that two k-letter words are in the same *equivalence class* if one is a cyclic permutation of the other. An equivalence class is called *complete* if it contains k distinct members. Comma-freedom is violated if we choose words from incomplete equivalence classes, or if more than one word is chosen from the same complete class. Thus, if we require that the words be formed from a fixed n-letter alphabet, we obtain an upper bound on the size of a comma-free code by counting the complete classes. Letting $W_k(n)$ denote the greatest number of words that such a dictionary can contain, we have **(2)**

$$W_k(n) \leqslant \frac{1}{k} \sum_{d|k} \mu(d) n^{k/d}, \tag{1}$$

where the summation is extended over all divisors d of k, and μ is the Möbius function defined by

$$\mu(d) = \begin{cases} 1 & \text{if } d = 1, \\ 0 & \text{if } d \text{ has any square factor,} \\ (-1)^r & \text{if } d = p_1p_2 \ldots p_r, \text{ where } p_1, \ldots, p_r \text{ are distinct primes.} \end{cases}$$

The first (last) n letters of a word are called an *initial* (*final*) *n-gram* (*digram*, *trigram*, *tetragram* being used for 2, 3, 4 respectively).

Golomb *et al.* **(2, 3)** investigate these codes and demonstrate many of their properties. In addition to this, Golomb, Welch, and Delbrück **(3)** and Golomb **(1)** provide an introduction to the biological aspects and applications of these codes. References to the biological literature are contained therein. Jaynes **(4)** cites the relevant engineering literature. The purpose of this paper is to report further results on the subject, obtained since the previous articles were written by a number of investigators, both human and electronic. Table I summarizes known values of $W_k(n)$ to date; entries not mentioned in **(2, 3)** are underlined. The arrows indicate that a row or column is known out to ∞. In § 6 we discuss a generalization *comma-free codes of index r*.

2. Results for even k. First we mention an improvement of Theorem 4 **(2)** which is due to Robert Jewett.

Received January 15, 1962. This research was supported in part by the National Aeronautics and Space Administration.

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

TABLE I.

$W_k(n)$

$k \diagdown n$	1	2	3	4	5	6	7	. . .	α_k
1	1	2	3	4	5	6	7	→	1
2	0	1	3	5	8	12	16	→	1/3
3	0	2	8	20	40	70	112	→	1/3
4	0	3	18	57					
5	0	6	48	204	624	1,554	3,360	→	1/5
6	0	9							
7	0	18	312	2,340	11,160	39,990	117,648	→	1/7
8	0	30							
9	0	56	2,184	29,120	217,000	1,119,720	4,483,696	→	1/9
10	0								
11	0	186	16,104	381,300	14,438,920	32,981,550	179,756,976	→	1/11
12	0								
13	0	630	122,640	5,162,220	93,900,240	1,004,668,770	7,453,000,800	→	1/13
14	0								
15	0	2,182	956,576	71,582,716	2,034,504,992	31,345,665,106	316,504,099,520	→	1/15
16	0								
17	0	7,710	7,596,480	1,010,580,540	44,878,791,360	995,685,849,690	13,684,147,881,600	→	1/17
	↓								

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

THEOREM 1. *If $k = 2i$, then the upper bound given by equation* (1) *is not attained by $W_k(n)$ provided that $n > 2^i + i$.*

Proof. Letting $0, \ldots, n-1$ be the letters of our alphabet, consider the equivalence classes containing words that have 0 in every position except possibly positions r and $r + i$, $1 \leqslant r \leqslant i$. For each letter d we define a sequence $x_1^d x_2^d \ldots x_i^d$ determined by the representatives of these classes, where

$$x_r^d = \begin{cases} 3 & \text{if } d \text{ appears in some representative in position } r, \text{ and in another in position } r+i, \\ 2 & \text{if } d \text{ appears in some representative in position } r+i, \text{ but never in position } r, \\ 1 & \text{if } d \text{ appears in some representative in position } r, \text{ but never in position } r+i, \\ 0 & \text{otherwise.} \end{cases}$$

First, we notice that if $d \neq b$ then x_r^d and x_r^b cannot both have the value 3. Since if they did we would have words of the form $w_1 = 0 \ldots 0d0 \ldots 0p0 \ldots 0$, $w_2 = 0 \ldots 0q0 \ldots 0d0 \ldots 0$, $w_3 = 0 \ldots 0b0 \ldots 0s0 \ldots 0$, $w_4 = 0 \ldots 0t0 \ldots 0b0 \ldots 0$. But then w_2w_3 and w_4w_1 would contain all possible representatives of the class $(0 \ldots 0d0 \ldots 0b0 \ldots 0)$ as overlaps. Secondly, if $d \neq b$, then there exists an r such that $0 \neq x_r^d \neq x_r^b \neq 0$. For if d is in the rth position of the representative of $(0 \ldots 0d0 \ldots 0b0 \ldots 0)$, then $x_r^d = x_r^b$ implies that $x_r^d = x_r^b = 3$, which is a contradiction. Further, for this r, x_r^d and x_r^b are not zero. Thus the maximum number of distinct sequences $\{x_r^d\}$ is the maximum number of distinct letters in a comma-free dictionary. Now the maximum number of distinct sequences containing a 3 is i. Further, we may replace every 0 in our sequences by 1, leaving them all distinct. Thus we may have 2^i distinct sequences not containing a 3.

THEOREM 2 (L. D. Baumert). *If D achieves the upper bound* (1) *for $k = 2i$, $n \geqslant 4$, then every letter occurs in every position.*

Proof. Suppose A does not occur in the jth position. Consider the complete equivalence classes containing words that have A in every position, except possibly positions j and $j + i$ (modulo k). There is a 1–1 correspondence between these classes and the complete classes for $k = 2$, same n. The eligible members of each such class for $k = 2i$ correspond to the eligible members of the corresponding class for $k = 2$. Further, this correspondence preserves comma-freedom. But codes for $k = 2$, $n \geqslant 4$ do not achieve the upper bound.

COROLLARY. *If $k = 2i$, $n \geqslant 4$, and D achieves the upper bound, then no initial $(k-1)$-gram is a final $(k-1)$-gram.*

Proof. If $a_1a_2 \ldots a_k$ and $ba_1a_2 \ldots a_{k-1}$ were in D, then a_k could not be initial, which contradicts Theorem 2.

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

THEOREM 3 (Alfred Hales). *If $k = 4$, $n > 2$, then D achieves the upper bound only if no initial trigram is a final trigram.*

Proof. Suppose $ABCD$ and $BCDE$ were in D; then $A \neq E$, and letting Z be a third letter, we see that $AAAE$, $AAAZ$, $ZEEE$, $AAEE$ must represent their classes. This means that the class containing $AAEZ$ cannot be represented.

THEOREM 4 (L. D. Baumert). $W_4(4) < 60$.

The proof consists of several lemmas. In order to condense these proofs the following conventions are assumed throughout:

(a) A, B, C, D are the letters of the alphabet.

(b) $(ACAD)$ denotes the equivalence class containing the word $ACAD$.

(c) $*(ABCB)*$ means that no element of $(ABCB)$ may be added to the partially formed dictionary without destroying comma-freedom.

(d) A sequence $ABCD$, $DCBB$, $*(CDBB)*$ means that because of overlap properties, existing hypotheses, and the Corollary to Theorem 2, the choice of $ABCD$ requires that $DCBB$ be chosen and these together imply $*(CDBB)*$.

(e) $|D|$ is the number of words in D.

LEMMA 1. *If $|D| = 60$ $(n = k = 4)$, every diagram occurs initially or finally.*

Proof. There are two types of diagrams (xx and xy). Let $x = D$, $y = C$; if DD does not occur, then $*(DDDC)*$. Suppose CD does not occur, then we have four cases: (1) $DDDC$, $CCDD$, $DBDC$, $DBDD$, $*(DBCC)*$; (2) $DCDD$, $DDCC$, $*(DCCC)*$; (3) $DCDD$, $CCDD$, $DCCC$, $ACDD$, $DCAC$, $*(CCCA)*$; (4) $DCDD$, $CCDD$, $CCDC$, $BCDD$, $CCDB$, $BCDC$, $DCDB$, $CCCB$, $BDDD$, $DCCB$, $BDDC$, $*(DBCB)*$.

LEMMA 2. *If $|D| = 60$ $(n = k = 4)$, then every letter begins (ends) at least two initial and two final digrams.*

Proof. Suppose that the only final digram beginning in A were AA (one exists by Theorem 2). Then AB, AC, AD are initial only (Lemma 1). Considering $(BACA)$, $(BADA)$, $(CADA)$, we see that one of BA, CA, DA is initial only, another final only, and the third both initial and final. From symmetry considerations, we need only investigate the case where BA is both initial and final, CA is initial only, DA is final only. Then $CADA$, $CABA$, $BADA$, $AADA$, $AABA$, but $AABA$ and $BADA$ with AA as a final digram contradict comma-freedom. Suppose the only final digram beginning in A were AB, then $AACA$, $AADA$, and considering $(CADA)$ we see that one of CA, DA is both initial and final, the other final only. Thus by symmetry we need only consider CA final only, DA both initial and final. Then $AACA$, $AADA$, $DACA$, $BACA$, $BADA$, $AADB$, $DADB$, $*(DBCA)*$. Symmetry takes

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

care of AC (or AD) as the only final digram. Proofs for initial digrams and for final letters are quite similar.

LEMMA 3. *If* $|D| = 60$ $(n = k = 4)$, *not every digram beginning (ending) in* X *is both initial and final.*

Proof. Let $X = A$ and suppose that every digram beginning (ending) in A were both initial and final. Considering $(AAAB)$, $(AAAC)$, $(AAAD)$, $(ABAC)$, $(ABAD)$, $(ACAD)$ we see that we cannot pick a comma-free set from them.

LEMMA 4. *If* $|D| = 60$ $(n = k = 4)$, *every letter begins a digram that is both initial and final.*

Proof. Let AA, AB be initial only; AC, AD final only (Lemma 2). Then $AABA$ and considering $(CABA)$, $(CADA)$, $(DABA)$ we see that one of CA, BA, DA is both initial and final, another final only, and the third initial only. Three cases arise: (1) CA both initial and final, DA initial, BA final; then $AADB$, $*(DAAA)*$. (2) CA both initial and final, DA final, BA initial; then $CADA$, $*(BACA)*$. (3) BA both initial and final, CA initial, DA final; then $AAAC$, $*(DAAC)*$. Assuming AA, AB are final only and AC, AD initial only, we have $BAAA$, $BBAA$, $CBAA$, $DBAA$ and (1) $BABD$, $DAAB$, $ADDA$, $DABD$, $*(DABA)*$ or (2) $BDBA$, $DBDA$, $DAAB$, $ADAB$, $BBDA$, $*(DBBB)*$.

LEMMA 5. *If* $|D| = 60$ $(n = k = 4)$, *no letter begins three digrams that are both initial and final.*

Proof. Let AA, AB, AC be such digrams. Considering $(AAAB)$, $(AAAC)$, $(ABAC)$, we have only two symmetric possibilities, one of which is $AABA$, $CAAA$, $CABA$, $AABC$, $CBAB$, $CBAA$, $ACBC$, $CBAD$. By Lemma 3, AD is final only, giving $DCBC$, $CAAD$, $CADB$, $*(DBAA)*$. Let AB, AC, AD be the digrams, then any comma-free solution for $(ABAC)$, $(ACAD)$, $(ADAB)$ leaves one of BA, CA, DA initial only, another final only, and the third both initial and final. Let BA be both initial and final, CA initial, DA final, then $CADA$, $BADA$, $CABA$, and by Lemma 3: (1) AA final only, $*(DAAA)*$; (2) AA initial only, $AAAC$, $ACDA$, $*(CBAA)*$.

LEMMA 6. *If* $|D| = 60$ $(n = k = 4)$, *no letter begins three digrams that are not both initial and final.*

Proof. Assume that A begins three such digrams. Then we have several cases to consider.

(I) AB, AC, AD are the digrams, AA is both initial and final. Then (1) AB, AC initial only, AD final only, by symmetry we may choose $CABA$, then $AABA$, $DABA$, (a) $CAAA$, $AABC$, $*(CBAB)*$, (b) $ACAA$, $AAAD$, $DBAA$, $*(DBAC)*$, (c) $ACAA$, $DAAA$, $AABD$, $*(DBAB)*$ and (2) AD initial only, AB, AC final only, choose $CABA$ as above, then (a) $AAAB$, $AADA$, $AAAC$, $BDAA$, $*(DBDA)*$, (b) $AAAB$, $ADAA$, $AAAC$, $BDAA$,

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

$CDAA$, $DADB$, $AADB$, $BDBA$, $DCDB$, $*(DBAC)*$, (c) $AABA$, $AABD$, $DBAB$, $ADDB$, $*(DBAA)*$, (d) $AABA$, $BDAA$, $DADB$, $AADB$, $*(DBBA)*$.

(II) One of AB, AC, AD is both initial and final. (1) AA initial only, AD, AC final only, AB both, we can choose $DACA$ as above (a) $AABC$, $CABA$, $AABD$, $AABA$, $AADB$, $AADC$, $CABD$, $*(DBAB)*$, (b) $AABC$, $BACA$, $AABD$, $AABA$, $AADC$, $AADD$, $BDCA$, $DDCA$, $DBAD$, $DBDD$, $BDCD$, $*(DCAC)*$, (c) $ABCA$, $AADC$, $AADD$, $AADA$, $AABD$, $DABD$, $*(DCAB)*$. (2) AA, AB initial only, AD final only, AC both, then $AABA$, (a) $AACA$, $CABA$, $AACC$, $CACC$, $CBAC$, $*(CBAB)*$, (b) $AACA$, $BACA$, $DACA$, $AADC$, $DABA$, $*(DAAA)*$, (c) $AAAC$, $ACBA$, $AADB$, $DABA$, $DAAC$, $*(DAAA)*$. (3) AA, AD final only, AB initial only, AC both, then $DAAA$, $DAAD$, $DAAC$, $*(DAAB)*$. (4) AB, AC initial only, AA final only, AD both, we can choose $CABA$ and then $BAAA$, $*(DBAA)*$.

LEMMA 7. *If all digrams beginning in a fixed letter are initial (final), then* $|D| < 60$ $(n = k = 4)$.

Proof. There are four cases. (1) AB, AC, AD, AA initial, AA, AB final, we can choose $DACA$ and then $AACA$, $BACA$, $BAAA$, $AADB$, $*(DBCA)*$. (2) AA, AB, AC, AD initial, AB, AC final, then $AADA$, we can choose $CABA$ and thus $CADA$, (a) $DABA$, $AAAC$, $*(DAAC)*$, (b) $BADA$, $AAAC$, $*(CBAA)*$. (3) AA, AD initial, AA, AB, AC, AD final, we can choose $CABA$, and then (a) $DAAA$, $CAAA$, $AABA$, $AABD$, $AABC$, $DBAB$, $DBAD$, $DBCB$, $*(DBCA)*$, (b) $AADA$, $CADA$, $BADA$, $AAAB$, $BDAA$, $*(DCAB)*$. (4) AB, AC initial, AA, AB, AC, AD final, we can choose $CABA$ and then $DAAA$, $DAAB$, $BAAA$, $*(BBAA)*$.

LEMMA 8. *If exactly three initial and three final digrams begin with the same letter, then* $|D| < 60$ $(n = k = 4)$.

Proof. There are three cases. (1) AA, AB both, AC initial, AD final (a) $BAAA$, $AACB$, $BADC$, $DACB$, $BADA$, $*(DAAA)*$, (b) $AABA$, $AABD$, $DABA$, $DBAB$, $*(DBAC)*$, (c) $AABA$, $BDAA$, $DABA$, $AAAD$, $DDAA$, $*(DDAB)*$. (2) AB, AC both, AA initial, AD final, we can choose $CABA$, (a) $AACC$, $AACA$, $CACC$, $CBAC$, $*(CBAA)*$, (b) $CAAC$, $ACBA$, $DABA$, $DAAC$, $AAAD$, $*(DBAA)*$. (3) AB, AC both, AA final, AD initial, we can choose $CABA$ and then $BAAA$, $*(BBAA)*$.

Proof of Theorem 4. Consider the digram structure for the letter A. All 25 possible cases are prohibited by Theorem 2 or one of the above lemmas.

Further, we have:

THEOREM 5. *If* $n = k = 4$, $|D| > 56$ *implies that no initial trigram is a final trigram.*

Proof. $xyzw$ and $yzwv$ in D imply (let $x = A, v = B$) that $(AAAB)$, $(AABB)$, $(ABBB)$, $(AAAD)$, $(AADB)$, $(ADBB)$, $(DBBB)$, $(AAAC)$, $(AACB)$, $(ACBB)$, $(CBBB)$ are represented by the elements indicated, if at all. $(DBAB)$,

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

(*CBAB*), (*CBDB*), (*ADAB*), (*DBAD*), (*DBAC*), (*ABAC*), (*ADAC*), (*ADCB*), (*ACCB*) must be represented by one of two elements, each of which is an overlap of the first set. At least four classes must be dropped to eliminate these overlaps.

3. Computational results. Robert Jewett was the first to exhibit a code for $n = 2$, $k = 8$, achieving the upper bound 30. Several more of these have been computed using an IBM-704 computer. One of these satisfies the inequalities:

$$\begin{array}{l} a \geqslant b < c \geqslant d \geqslant e, \quad f \geqslant g \geqslant h \\ a < b \geqslant c < d \geqslant e < f \geqslant g \geqslant h \\ a \geqslant b < c \geqslant d = e = f < g = h \\ a > b < c = d \geqslant e = f < g = h \end{array}$$

An exhaustive search was made to determine $W_4(4)$, the results of which show that $W_4(4) = 57$. Since this was done by digital computer and is quite time-consuming, it cannot be verified by hand. Thus a proof of this fact without using a computer would still be of some interest. To this end it is interesting to note that in all 57 word dictionaries computed (*not* all 57's were computed), classes of the same kind were missing. Specifically, each code had one class of the type (*xyzw*) and two classes of the type (*pqpr*) missing.

A typical code achieving 57 is the following:

DBDD
DCCB
CACB
BCAD
BDAD
ACCD
BBAD
DBCD
DBBB
BABD
BACD
ACAA

DDAD
ACCA
DACB
BAAD
BDCD
DABB
DCAB
BAAB
CCCA
BBCB
DCAA
ABCD

DDCD
DABD
DCAD
ACAB
CCCD
BCAB
CACD
BCAA
BCCA
BACB
ABAD

CCCB
BABB
DACD
BAAA
BCCD
ABCB
DBAD
DCBB
DCCA
AACD
DAAA

BCCB
DBCB
DAAD
DCBD
DCCD
DAAB
ACAD
DBBD
ACCB
BBCD
AACB

After considerable computation, it appeared that an exhaustive search for codes achieving the upper bound (=116) in the case of $n = 3$, $k = 6$, would not be feasible. However, it was established by the computer that 116 cannot be achieved if *AABBCC* is used to represent its class. Thus, any 116-code is isomorphic to one containing *ABBCCA*. The computer program which produced these results was written by Lee Laxdal.

4. Results for odd *k*. First we mention that a slight error occurs in the proof of Theorem 2 **(2)**. Specifically there are 9 (rather than 8) patterns of

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

proper differences of $k = 7$. The missing difference pattern is $(+ - + - + -)$. The result is still valid, however. Using this method of proof it has been established that $17W_{17}(n) = n^{17} - n$.

Further, John Selfridge has shown that for all odd k, the upper bound (1) is attained for a class of dictionaries satisfying a weaker constraint than comma-freedom.

DEFINITION. *A collection D of k-letter words over an n-letter alphabet will be called locally decipherable if, whenever $(a_1a_2 \ldots a_k)$ and $(b_1b_2 \ldots b_k)$ are in D, the following three conditions do not hold simultaneously for any $i = 1, 2, \ldots, k-1$:*

(i) *$a_1a_2 \ldots a_i$ is a final i-gram in D,*

(ii) *$a_{i+1} \ldots a_kb_1 \ldots b_i$ is a word of D,*

(iii) *$b_{i+1}b_{i+2} \ldots b_k$ is an initial $(k-i)$-gram in D.*

It is clear that every comma-free dictionary is locally decipherable, but not necessarily conversely.

THEOREM 6. *Letting $X_k(n)$ be the maximum size for a locally decipherable dictionary of k-letter words (k odd) over an n-letter alphabet, we have*

$$X_k(n) = \frac{1}{k} \sum_{d \mid k} \mu(d) n^{k/d}.$$

Dr. Selfridge's result casts reasonable doubt on whether $W_k(n)$ must achieve the upper bound (1) for all odd k, as was conjectured previously **(2)**.

5. Asymptotics. An improved asymptotic bound for

$$\alpha_k = \lim_{n \to \infty} \frac{W_k(n)}{n^k}, \qquad k \text{ even},$$

has been worked out by Basil Gordon. (In **(2)** it was shown that α_k exists for all k, that $\alpha_k = 1/k$ for odd k, that $\alpha_2 = 1/3$, and that $1/ek < \alpha_k \leqslant 1/k$ for even $k > 2$.)

THEOREM 7. *For even $k > 4$, $2/ek < \alpha_k \leqslant 1/k$, where $e = 2.71828\ldots$.*

Proof. We construct a comma-free code D containing

$$n^k \frac{2}{k}\left(1 - \frac{2}{k}\right)^{\frac{1}{2}k-1}$$

words, thus showing that

$$\alpha_k \geqslant \frac{2}{k}\left(1 - \frac{2}{k}\right)^{\frac{1}{2}k-1} > \frac{2}{ek}.$$

First form a comma-free dictionary D' of two-letter words from our n-letter alphabet containing $[n^2/h]$ words, where $2h = k$. This is possible if $h > 2$, that is, if $k > 4$ **(2)**. Partition the set X of ordered pairs of letters from our

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

n-letter alphabet into two disjoint sets, D' and $X - D'$. Let a_1b_1 be an arbitrary pair in D', and similarly a_ib_i $(i \neq 1)$ an arbitrary pair in $X - D'$. Now the comma-free code we seek is composed of all words of the type $a_1a_2 \ldots a_h$ $b_1b_2 \ldots b_h$.

Applying this construction to $n = 2$, $k = 6$, the set D' consists of the single word 01 and $X - D' = \{00, 10, 11\}$, giving rise to code 000100, 001100, 001101, 010100, 011100, 011101, 010110, 011110, 011111, which is maximal for these parameters. For $n = 3$, $k = 6$ this construction gives a code of size 108 (the bound here is 116). For $n = 5$, $k = 4$ we get 136 and the bound is 150.

6. Comma-free codes of index r.

The results of this section are due to S. W. Golomb.

A set D of k-letter words is called a *comma-free code of index r* if whenever $(a_1a_2 \ldots a_k)$ and $(b_1b_2 \ldots b_k)$ are in D, then each of the overlaps $(a_2a_3 \ldots a_kb_1)$, $(a_3a_4 \ldots a_kb_1b_2), \ldots, (a_kb_1b_2 \ldots b_{k-1})$ differs from every word of D in at least r places. In this terminology ordinary comma-free codes are of index 1. Letting $W_r(n, k)$ denote the largest number of k-letter words that a comma-free code of index r can have over an n-letter alphabet, then Theorems 8 to 11 are immediate consequences of the definition.

THEOREM 8. $W_0(n, k) = n^k$,

THEOREM 9. $W_r(n, k) = 0$ for $r = 0, 1, 2, \ldots$.

THEOREM 10. $W_r(n, k) = 0$ for $r > k > 1$.

THEOREM 11. *$W_r(n, k)$ is a monotonic non-decreasing function of n; $W_r(n, k)$ is a monotonic non-increasing function of r.*

THEOREM 12. *$W_r(n, 2) = n^2$ for $r = 0$, $[n^2/3]$ for $r = 1$, $[n^2/4]$ for $r = 2$, and 0 for $r > 2$.*

Proof. Theorems 8 and 10 take care of $r = 0$ and $r > 2$ respectively. Theorem 3 of **(1)** provides $[n^2/3]$ for $r = 1$. If $r = 2$, we note that no letter occurs both initially and finally in such a code. Thus the best we can do is divide the letters up evenly. This gives $[n^2/4]$.

THEOREM 13. *$W_r(n, k) = 1$ if $r = k = n > 0$.*

Proof. Note that every word must contain all $n = k = r$ letters. Now if $(a_1a_2 \ldots a_k)$, $(b_1b_2 \ldots b_k)$ are two such words, there is a first $b_i \neq a_i$. But $b_i = a_j$ for some $j > i$, say for $j = i + h$. Then the overlap $a_{1+h} \ldots a_na_1 \ldots a_h$ agrees with $b_1 \ldots b_n$ at b_i. Thus $W_r(n, k) = 1$.

THEOREM 14. $W_2(n, 3) \geqslant \max\{[\tfrac{1}{3}n](n - [\tfrac{1}{3}n])^2, [\tfrac{1}{3}(n+3)](n - [\tfrac{1}{3}(n+3)])^2\}$.

Proof. Divide the alphabet into two disjoint sets A, B where $|A| \leqslant |B|$. Consider all words of the form abb, clearly comma-free of index 2. In

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press

order to maximize the size of this code we seek the relative maximum of $f(x) = x(n-x)^2$ in the range $0 < x < n$. This occurs at $n/3$, giving the above result.

THEOREM 15. *$W_T(n, r) = q^{r-d}(q+1)^d$, where $q = [n/r]$, $d = n - qr$.*

Proof. All overlaps must differ from all words in every position; thus the sets of letters occurring in positions $1, 2, \ldots, r$ are disjoint. Maximizing the size of such a code gives the result.

REFERENCES

1. S. W. Golomb, *Proceedings of the Symposium on Mathematical Problems in the Biological Sciences*, 1961, Amer. Math. Soc. (to appear).
2. S. W. Golomb, Basil Gordon, and L. R. Welch, *Comma-free codes*, Can. J. Math., *10* (1958), 202–209.
3. S. W. Golomb, L. R. Welch, and M. Delbrück, *Construction and properties of comma-free codes*, Biol. Medd. Dan. Vid. Selsk., *23*, no. 9 (1958).
4. E. T. Jaynes, *Note on unique decipherability*, IRE Transactions on Information Theory (Sept. 1959), 98–102.

California Institute of Technology

https://doi.org/10.4153/CJM-1963-021-1 Published online by Cambridge University Press