Restrictions and Generalizations on Comma-Free Codes

Alexander L. Churchill

Student
Stanford University, California, USA

achur@stanford.edu

Submitted: Feb 13, 2008; Accepted: Feb 14, 2009; Published: Feb 20, 2009
Mathematics Subject Classifications: 94B50, 94B65

Abstract

A significant sector of coding theory is that of comma-free coding; that is, codes which can be received without the need of a letter used for word separation. The major difficulty is in finding bounds on the maximum number of comma-free words which can inhabit a dictionary. We introduce a new class called a self-reflective comma-free dictionary and prove a series of bounds on the size of such a dictionary based upon word length and alphabet size. We also introduce other new classes such as self-swappable comma-free codes and comma-free codes in q dimensions and prove preliminary bounds for these classes. Finally, we discuss the implications and applications of combining these original concepts, including their implications for the NP-complete Post Correspondence Problem.

1 Introduction

1.1 Comma-free codes

Comma-free codes were first introduced by Crick, Griffith, and Orgel [2] in 1957 as a potential explanation for the fact that DNA codes only twenty amino acids, despite the fact that it is a code with word-length three and a four-letter alphabet. While this explanation was revealed to be incorrect, comma-free codes are still a major area of exploration in coding theory. Initially, we establish definitions.

Let n be a fixed positive integer. Consider a dictionary of words in which each word has length k chosen from an n-letter alphabet. Let the alphabet consist of letters a_1, a_2, $a_3, \ldots, a_n$.

A set D of k-letter words is called a *Comma-Free Dictionary* (according to Golomb, Gordon, and Welch [4]) if whenever words $a_1a_2\cdots a_k$ and $b_1b_2\cdots b_k$ are in D, the "overlaps" $a_2a_3\cdots a_kb_1$, $a_3\cdots a_kb_1b_2$, $\ldots$, $a_kb_1b_2\cdots b_{k-1}$ are not in D.

The major problems investigated have been in determination of the maximum number of words a comma-free dictionary can possess, according to Levenshtein [6]. If the size of each word is k and the size of the alphabet is n, the maximum number of elements in D is denoted as $W(k,n)$. Golomb, Gordon, and Welch [4] established a bound for the maximum size of a comma-free dictionary as

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

where $\mu(d)$ is the Möbius function. This bound is established by noticing several phenomena.

Initially, we consider equivalence classes of words formed by taking cyclic shifts of the letters of that word. We have equivalence classes $\overline{\omega}$ which contains all cyclic shifts $\phi^i(\omega)$. We define a cyclic shift $\phi^i(\omega)$ where $\phi(a_1a_2\cdots a_k) = a_2a_3\cdots a_ka_1$. For instance, `ABCD` and `CDAB` are cyclic shifts of each other, so they are in the same equivalence class. Furthermore, we observe that a comma-free dictionary cannot contain more than one member from each equivalence class. To show this, consider the overlaps formed by repeating one word in the equivalence class. This yields overlaps of all other words in the equivalence class. Repeating `ABCD` gives `ABCDABCD` which contains `CDAB` as an overlap.

Golomb, Gordon, and Welch [4] also put forth the concept of subperiod. Let d be a divisor of k. We say that a word $a_1a_2\cdots a_k$ has subperiod d if it is of the form $a_1a_2\cdots a_d a_1a_2\cdots a_d \cdots\cdots a_1a_2a_d$. If a word has subperiod $d < k$, such as `ABCABC`, it cannot be contained in a comma-free dictionary, because repeating such a word to yield `ABCABCABCABC` contains the original word as an overlap. We call a word with subperiod $d = k$ *primitive*.

The bound (1) is calculated by counting all equivalence classes with subperiod k. Golomb, Gordon, and Welch [4] provedthis bound was tight for $k = 1$, 3, 5, 7, 9, 11, 13, and 15, and conjectured that it was tight for all odd k. This was proved by Eastman [3] in 1965. The only tight bound for even k was given by Golomb, Gordon, and Welch [4]. They found that

$$W(2,n) \leq \left\lfloor \frac{1}{3}n^2 \right\rfloor. \tag{2}$$

Finding a general tight bound for all even k is an open problem.

2 Self-reflective comma-free codes

One focus of this paper is Self-Reflective Comma-Free Codes. Initially, we must establish a definition. Let $\sigma(a_1a_2\cdots a_k) = a_ka_{k-1}\cdots a_2a_1$. We note that for every comma-free dictionary $D = \{\omega_1, \omega_2, \ldots, \omega_x\}$, there is a similar comma-free dictionary $D = \{\sigma(\omega_1), \sigma(\omega_2), \ldots, \sigma(\omega_x)\}$.

Definition: A set $D_r \subseteq D$ (where D is a comma-free dictionary) is called a self-reflective comma-free dictionary if for all words $\omega \in D_r$, $\sigma(\omega) \in D_r$. The focus of this paper is to establish bounds on the maximum size of self-reflective comma-free dictionaries for general n and k. Denote the greatest number of words D_r can possess as $W_r(k,n)$.

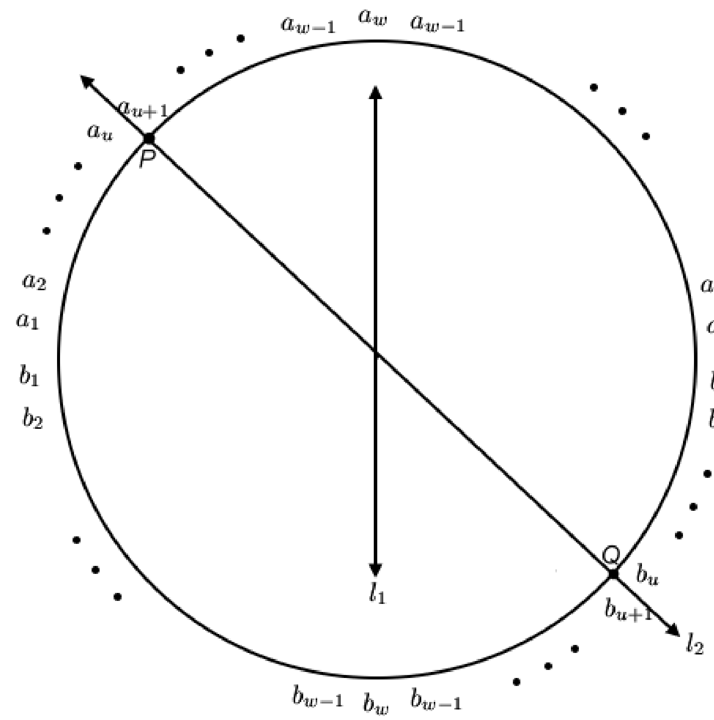

Figure 1: Bijective Circle

2.1 Results

2.1.1 Lemmas

We utilize the following lemmas for assistance in proving bounds on the size of self-reflective comma-free dictionaries. They give insight into word structure and properties of specific word types.

Lemma 1. *If $\sigma(\omega_1) \in \overline{\omega_1}$, then $\sigma(\phi^i(\omega_1)) \in \overline{\omega_1}$ for all i.*

Proof. When $i = 0$, the proof is trivial. Assume $i > 1$.

$$\sigma(a_{i+1}a_{i+2}\cdots a_k a_1 a_2 \cdots a_{i-1}a_i) = a_i a_{i-1}\cdots a_2 a_1 a_k \cdots a_{i+2}a_{i+1},$$

but we know $a_{i-1}a_{i-2}\cdots a_2 a_1 a_k \cdots a_{i+1}a_i \in \overline{\omega_1}$, so $a_i a_{i-1}\cdots a_2 a_1 a_k \cdots a_{i+2}a_{i+1} \in \overline{\omega_1}$.

This completes our proof. □

Lemma 2. *Let $\omega = a_1 a_2 \cdots a_{w-1}a_w a_{w-1}\cdots a_2 a_1 b_1 b_2 \cdots b_{w-1}b_w b_{w-1}\cdots b_2 b_1$. If ω is primitive, then there does not exist any ω_1 such that $\omega_1 \in \overline{\omega}$ and $\sigma(\omega_1) = \omega_1$.*

Proof. Assume some ω_1 exists. Let $\omega_1 = b_u b_{u-1}\cdots b_1 a_1 \cdots a_w \cdots a_1 b_1 \cdots b_w \cdots b_{u+1}$.

Consider a bijective circle in which each letter of ω_1 is represented by a coloring of points around a circle, as shown in Figure 1. This figure, by construction, is fixed under reflection about $l_1 = \overleftrightarrow{a_w b_w}$. Furthermore, we assume ω is self-reflective, so it must also

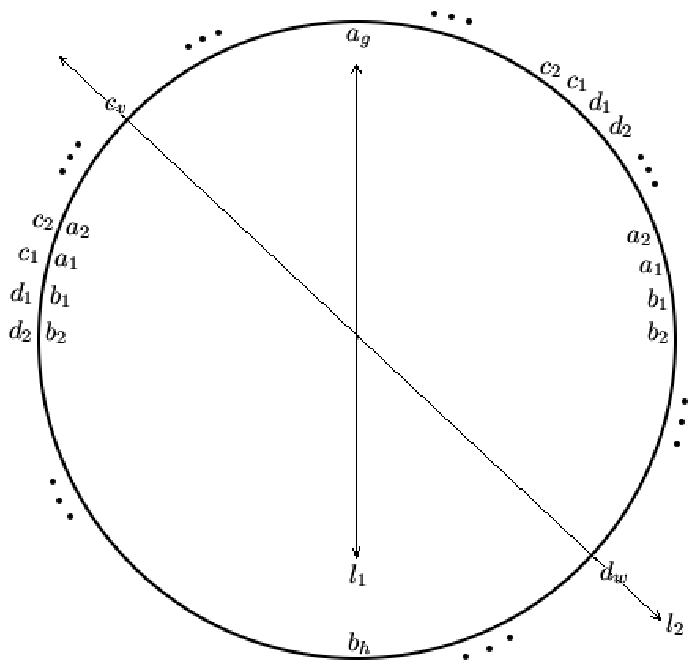

Figure 2: Bijective Circle

be fixed under reflection about $l_2 = \overleftrightarrow{PQ}$ where P and Q are the midpoints of $\overline{a_k a_{k+1}}$ and $\overline{b_k b_{k+1}}$ respectively.

But since the circle-word is fixed under reflection about l_1 and l_2, where $l_1 \neq l_2$, it is also fixed under the nonidentity rotation $l_1 \circ l_2$. Since it is fixed under some nonidentity rotation, the word itself must be fixed under some cyclic shift $\phi^i(\omega)$ where $i \neq k$. But since it is fixed under some such cyclic shift, it must have some subperiod such that $d|k$ and $d \neq k$. Thus it is not primitive. This contradiction proves the lemma. $\square$

Lemma 3. *Every word ω such that $\sigma(\omega) \in \overline{\omega}$ takes the form $\omega_1\omega_2$ where ω_1 and ω_2 are palindromes. Call such a word doubly palindromic.*

Proof. Assume without loss of generality that $\omega = a_1a_2 \cdots a_{k-1}a_k$ and let
$\sigma(\omega) = a_u a_{u+1} \cdots a_{k-1}a_k a_1 a_2 \cdots a_{u-1}$.
But then $a_u a_{u+1} \cdots a_{k-1}a_k a_1 a_2 \cdots a_{u-2}a_{u-1} = a_k a_{k-1} \cdots a_{u+1}a_u a_{u-1}a_{u-2} \cdots a_2 a_1$.
Clearly $a_u a_{u+1} \cdots a_{k-1}a_k$ and $a_1 a_2 \cdots a_{u-2}a_{u-1}$ are palindromes.
Thus, the word takes the desired form, which completes our proof. $\square$

Lemma 4. *If $\omega_1 = \omega_2$ where $\omega_1 = a_1a_2a_3 \cdots a_g \cdots a_3a_2a_1b_1b_2b_3 \cdots b_h \cdots b_3b_2b_1$ and $\omega_2 = c_1c_2c_3 \cdots c_v \cdots c_3c_2c_1d_1d_2d_3 \cdots d_w \cdots d_3d_2d_1$, then ω_1 and ω_2 have subperiod of length $\gcd(|i-j|, k)$ where $i = 2g-1$ and $j = 2v-1$.*

Proof. Consider a bijective circle as in Lemma 2, shown in Figure 2.

By construction, both words ω_1 and ω_2 are fixed under reflection about $l_1 = \overline{a_g b_n}$ and $l_2 = \overline{c_v d_w}$, so they are fixed about the rotation $l_1 \circ l_2$ which rotates each letter by twice the angle of the intersection of l_1 and l_2. That is, each letter rotates by $2(g-v) = i-j$. Thus any two letters separated by $i-j$ will be equal. This rotation generates the same subgroup of D_k as does rotation by $\gcd(|i-j|, k)$. Therefore the subperiod is of the desired length. □

2.1.2 Results for specific k

Theorem 1. *$W_r(2,n) = 0$ for all n*

Proof. We prove by contradiction. Assume $W_r(2,n) > 0$. Let $F_n = a_1 a_2 \cdots a_n$ be an n-letter alphabet. Suppose there exists a word in our dictionary, D_r.

Without loss of generality, $\omega_1 \in D_r$ where $\omega_1 = a_1 a_2$. Then $\sigma(\omega_1) \in D_r$ so $a_2 a_1 \in D_r$.

But $a_2 a_1$ is a cyclic shift of $a_1 a_2$ which cannot be part of a comma-free dictionary according to Crick, Griffith and Orgel [2]. This is a contradiction which completes our proof. □

Theorem 2. *$W_r(3,n) \leq \frac{2n^3-3n^2+n}{6}$;*

Proof. We use bound (1) which counts the number of equivalence classes with subperiod k

$$W(k,n) \leq \frac{1}{k} \sum_{d|k} \mu(d) n^{k/d}.$$

This gives us $W(3,n) \leq \frac{1}{3}(n^3 - n)$.

But this includes the equivalence classes $\overline{abb}$ and $\overline{aba}$. We cannot have both `aba` and `bab` in our comma-free dictionary, so for each pair of letters, there is either a counted word of the form `abb` or `bba` or of `aab` or `baa`. Without loss of generality, assume we have `abb` and `bba`. (In a self-reflective dictionary, both or neither must appear.) Since they are members of the same equivalence class, neither can appear, so we can subtract the equivalence class from our upper bound. There is one such equivalence class for every two letters which we can eliminate, for a total of $\binom{n}{2}$ total. We subtract to get

$$W_r(3,n) \leq \frac{2n^3 - 3n^2 + n}{6}.$$

□

Theorem 3. *$W_r(3,n) = \frac{2n^3-3n^2+n}{6}$.*

Proof. We use the construction given by Crick, Griffith, and Orgel [2] for n letters, removing those of the form ABB. Use the numbers 1 through n to represent an n-letter comma-free alphabet, giving a well-ordered set. In this description, $AB \begin{matrix} A \\ B \end{matrix}$ represents ABA and ABB.

$$
1\;2\;1 \qquad \begin{matrix}1\\2\end{matrix}\;3\;\begin{matrix}1\\2\end{matrix} \qquad \cdots \qquad \begin{matrix}1\\2\\3\\ \cdot\\ \cdot\\ \cdot\\ n-2\\ n-1\end{matrix}\;\; n \;\;\begin{matrix}1\\2\\3\\ \cdot\\ \cdot\\ \cdot\\ n-2\\ n-1\end{matrix}
$$

This is a comma-free code which has $1^2 + 2^2 + 3^2 + \cdots n^2 = \frac{2n^3+3n^2+n}{6}$ members. It is also self-reflective, because for all words abc, cba must also be a member. This proves the bound from Theorem 2 is tight. $\square$

2.1.3 Results for k odd

Theorem 4. *For odd k, $W_r(k,n) \leq \frac{1}{k}\sum_{d|k} \mu(d)n^{k/d} - \binom{n}{2}$.*

Proof. Consider equivalence classes $\overline{ababab\cdots aba}$ and $\overline{bbababa\cdots ba}$. Take words ω_1 and ω_2 in our dictionary from each respective equivalence class. Both $\sigma(\omega_1) = \omega_1$ and $\sigma(\omega_2) = \omega_2$ cannot be true. This is because then both $\mathtt{abab}\cdots\mathtt{aba}$ and $\mathtt{baba}\cdots\mathtt{bab}$ would necessarily be ω_1 and ω_2. This is not comma-free, because $\mathtt{(abab}\cdots\mathtt{aba)(baba}\cdots\mathtt{bab)}$ would then have ω_1 and ω_2 as an overlap. Thus, at least one word from one of the two equivalence classes must not reflect to itself. However, a reflection of either one of the equivalence classes yields a cyclic shift of that equivalence class, which is not allowed in a comma-free dictionary. Thus we subtract at least one of these two equivalence classes from bound (1). We subtract an equivalence class for each two letters, so there are a total of $\binom{n}{2}$ eliminated, giving us our desired bound. $\square$

2.1.4 Results for k even

Theorem 5. *For $k = 2$ (mod 4),*

$$W_r(k,n) \leq \frac{1}{k}\sum_{d|k} \mu(d)n^{k/d} - \binom{n^{(k+2)/4}}{2} + \sum_{d|\frac{k}{2}, d\neq\frac{k}{2}} \binom{n^{(d+1)/2}}{2}.$$

Proof. Consider a word

$$\omega = a_1a_2\cdots a_{s-1}a_sa_{s-1}\cdots a_1b_1b_2\cdots b_{s-1}b_sb_{s-1}\cdots b_2b_1.$$

We call such a word fixed doubly palindromic. Now let $\omega_1 \in \overline{\omega}$. Since $\sigma(\omega) = \phi^{k/2}(\omega)$, by Lemma 1, all ω_1 will have property $\sigma(\omega_1) \in \overline{\omega}$. Furthermore, assume

$$a_1a_2\cdots a_{s-1}a_sa_{s-1}\cdots a_1a_1 \neq b_1b_2\cdots b_{s-1}b_sb_{s-1}\cdots b_2b_1.$$

Then ω and subsequently ω_1 cannot have an even subperiod. By Lemma 2, any such word which is a palindrome must have a subperiod. If a fixed doubly palindromic word is not a palindrome, we can remove its equivalence class from our bound, as reflection of that word would yield a nonidentity cyclic shift of that word. We count the number of non-palindromic classes by counting all fixed doubly palindromic classes and subtracting the fixed doubly palindromic classes with subperiod $d \neq k$. The number of fixed doubly palindromic equivalence classes is established by first counting the number of possible palindromes $a_1 a_2 \cdots a_{s-1} a_s a_{s-1} \cdots a_1$. We know $s = \frac{k+2}{4}$. Thus the number of such palindromes is $n^{(k+2)/4}$. We then choose two distinct such palindromes to form our equivalence class, giving the total number of equivalence classes as $\binom{n^{(k+2)/4}}{2}$. To count the number of equivalence classes with nontrivial subperiods, we first note that all odd subperiods of length d have the property that $d|\frac{k}{2}$. Furthermore, since the equivalence classes with subperiod we are counting form a palindrome, the subperiod word itself must be palindromic. Therefore, the number of possible different subperiods of length d is $\binom{n^{(d+1)/2}}{2}$. The total number of equivalence classes with subperiod, therefore, is $\displaystyle\sum_{d|k/2, d\neq \frac{k}{2}} \binom{n^{(d+1)/2}}{2}$.

Thus, the number of primitive equivalence classes of form ω is $\binom{n^{(k+2)/4}}{2} - \displaystyle\sum_{d|\frac{k}{2}, d\neq\frac{k}{2}} \binom{n^{(d+1)/2}}{2}$.

Subtracting from the original bound (1), we complete our proof. □

Theorem 6. *For k even:*

$$W_r(k,n) \leq \left(\frac{1}{k}\sum_{d|k}\mu(d)n^{k/d}\right) - \frac{kn^{(k+2)/2}}{4} + \sum_{i,j\leq\frac{k}{2},\ i,j\ \ odd} \frac{\gcd(|i-j|,k)n^{\frac{\gcd(|i-j|,k)+2}{2}}}{4}.$$

Proof. Consider a word $\omega = a_1 a_2 \cdots a_v \cdots a_2 a_1 b_1 b_2 \cdots b_w \cdots b_2 b_1$. Note that such a word is doubly palindromic. Clearly $\sigma(\omega) \in \overline{\omega}$. Now consider the equivalence class $\overline{\omega}$. We begin by counting those equivalence classes. We initially observe $v + w = \frac{k+2}{2}$. There are a total of $\frac{k}{2}$ possible values for v (and subsequently w), since the length of both palindromes $a_1 a_2 \cdots a_v \cdots a_2 a_1$ and $b_1 b_2 \cdots b_w \cdots b_2 b_1$ must be odd. This gives $kn^{(k+2)/2}$. However, this will count both ω and $\phi^{2v-1}(\omega)$. Therefore, we divide by two to find our total number of equivalence classes. Thus the total number of such equivalence classes is $\frac{kn^{(k+2)/2}}{4}$.

However if $\omega_1 = \omega_2$, where

$$\omega_1 = a_1 a_2 a_3 \cdots a_g \cdots a_3 a_2 a_1 b_1 b_2 b_3 \cdots b_h \cdots b_3 b_2 b_1$$

$$\omega_2 = c_1 c_2 c_3 \cdots c_v \cdots c_3 c_2 c_1 d_1 d_2 d_3 \cdots d_w \cdots d_3 d_2 d_1,$$

there is overcounting. By Lemma 4, such a situation forces ω_1 and ω_2 to have a subperiod of length $\gcd(|i-j|,k)$ where $i = 2g-1$ and $j = 2v-1$. To count these equivalence classes, we assume without loss of generality that each i and j is at most $\frac{k}{2}$. Furthermore, we note that since we have a word such that $\sigma(\omega_1) \in \overline{\omega}_1$, the subperiod must have the same property. By Lemma 3, this means the subperiod must take the form $g_1g_2\cdots g_r\cdots g_2g_1h_1h_2\cdots h_t\cdots h_2h_1$. We proceed to count all such subperiods using a method similar to that used to count all doubly palindromic words. This yields $\sum_{i,j\leq\frac{k}{2}}\frac{\gcd(|i-j|,k)n^{\frac{\gcd(|i-j|,k)+2}{2}}}{4}$. Furthermore, by Lemma 4, this also counts the total number of words with subperiod in our original count.

We subtract to yield $\frac{kn^{(k+2)/2}}{4}-\sum_{i,j\leq\frac{k}{2}}\frac{\gcd(|i-j|,k)n^{\frac{\gcd(|i-j|,k)+2}{2}}}{4}$ as the total number of doubly palindromic equivalence classes without subperiods or overcounts. Since each of these classes produces a word whose reflection is also a cyclic shift, none can be contained in a self-reflective comma-free dictionary. Thus we can subtract this number ω from the original bound (1) to gain our desired result. $\square$

2.2 Applications

Despite the youth of self-reflective comma-free codes many applications have surfaced. The problem which inspired self-reflective coding is that of efficient use of a receiver. The receiver needs to know fewer words, as it can compare both a string of letters and the reflection of that string to synchronize the code. This is especially useful when a receiver needs to be particularly space-efficient. Furthermore, self-reflective comma-free codes can be used as bijections to a variety of palindromic problems. Apart from the obvious applications for combinatorial problems regarding palindromes, there are a variety of other ramifications. A tight bound on the size of a self-reflective comma-free dictionary when k is even would give a lower bound on the size of a standard comma-free dictionary for even k. This is particularly useful, because it bounds a quantity from below which is already bounded from above, and has ramifications for the applications of standard comma-free codes.

3 Self-swappable comma-free dictionaries

We define a dictionary D_s to be self-swappable if it is fixed under the permutation $f(\omega)=(a_1a_2)(a_3a_4)\cdots(a_{n-1}a_n)$ where all a_i are members of an n-letter alphabet where n is even. We denote the maximum number of words a self-swappable comma-free dictionary can contain given k-letter words and an n-letter alphabet as $W_s(k,n)$.

Lemma 5. *If $\omega\in D_s$ and $f(\omega)\in\overline{\omega}$, either $f(\omega)=\phi^{k/2}(\omega)$ or ω has subperiod $d\neq k$.*

Proof. We know the permutation $f(\omega)$ has order 2. Thus if $f(\omega)=\phi^m(\omega)$, then $\omega=\phi^{2m}(\omega)$. In other words, such a word must be fixed under a cyclic shift of size $2m$. It follows that either $k=2m$ or the word has some subperiod $d\neq k$ (as any word fixed under a nonidentity cyclic shift is not primitive). This observation completes the proof. $\square$

Theorem 7. *For n and k even,*

$$W_s(k,n) \leq \frac{1}{k}\sum_{d|k} \mu(d) n^{k/d} \; - \frac{1}{k}\left(n^{k/2} - \sum_{d|k,\ k/d\ odd} n^{d/2} \right)$$

Proof. To determine this bound, we remove the number of equivalence classes $\overline{\omega}$ satisfying $f(\omega) \in \overline{\omega}$ from bound (1). We remove these, because for all words $\omega \in D_s$, $f(\omega) \in D_s$. Since $f(\omega)$ is a cyclic shift of ω, we remove the equivalence class. We count the size of the equivalence class by first counting the number of words ω_1 which have the property that $f(\omega_1) = \phi^{k/2}(\omega_1)$. This number is found by constructing words $\omega_1 = a_1 a_2 a_3 \cdots a_{k/2} b_1 b_2 b_3 \cdots b_{k/2}$ where permutation f takes all a_i to all respective b_i. The number of such words is $n^{k/2}$. We then subtract the number of words ω_1 which have subperiod $d \neq k$. We know k/d cannot be even, because that would require all a_i and b_i be equal, which is never true. This means k/d is odd. Furthermore, since k/d is odd, the subperiod must take the form $a_{k/2-d} \cdots a_{k/2-1} a_{k/2} b_1 b_2 \cdots b_d$. Furthermore, the first half of the subperiod in this section must be the same as the first half of the subperiod starting the word. Thus the subperiod must take the form $a_1 a_2 \cdots a_d b_1 b_2 \cdots b_d$. This means we can count the subperiod by $\displaystyle\sum_{d|k,\ k/d\ odd} n^{d/2}$. We then subtract this from our count of all words of form ω_1 and divide by k to count the number of equivalence classes. Subtracting from the original inequality gives our desired bound. $\square$

3.1 A construction for self-swappable comma-free dictionaries of word-length three

We consider the original construction for dictionaries of word-length 3 given by Crick, Griffith, and Orgel [2]. We slightly modify this original construction to create a self-swappable dictionary. In this construction, $AB\begin{matrix}A\\B\end{matrix}$ represents ABA and ABB and the numbers 1 through n represent an n-letter alphabet.

$$\begin{array}{ccccccccccc}
\begin{matrix}1\\2\end{matrix} & \begin{matrix}3\\4\end{matrix} & \begin{matrix}1\\2\\3\\4\end{matrix} & \quad & \begin{matrix}1\\2\\3\\4\end{matrix} & \begin{matrix}5\\6\end{matrix} & \begin{matrix}1\\2\\3\\4\\5\\6\end{matrix} & \cdots & \begin{matrix}1\\2\\3\\4\\.\\.\\.\\n-3\\n-2\end{matrix} & \begin{matrix}n-1\\n\end{matrix} & \begin{matrix}1\\2\\3\\4\\.\\.\\.\\n-3\\n-2\\n-1\\n\end{matrix}
\end{array}$$

This construction is comma-free and self-swappable. It gives a total of $\frac{n^3-4n}{3}$ words over an n-letter dictionary. This differs by the bound for standard comma-free code dictionaries of size 3 by exactly n from bound (1) which for $k = 3$ is $\frac{n^3-n}{3}$. An improved construction or proof of tighter bound is an open problem.

4 Comma-free matrices and q-dimensional comma-free codes

Now consider a new type of problem in which we define a comma-free matrix dictionary D^2 as a set containing matrices with dimensions k_1 by k_2 which have the property that for any arrangement of matrices from D^2 on a plane, any "overlaps" are not in D^2. That is to say, any k_1 by k_2 array chosen in a plane of letters created by words from D^2 is not in D^2. We extend the problem to any q-dimensional array of letters. We denote a q-dimensional comma-free dictionary as D^q. The maximum number of words such a dictionary can contain over n letters and with word-size of $k_1 \times k_2 \times \cdots \times k_q$ is denoted as $Q(k_1, k_2, \ldots, k_q, n)$.

4.0.1 Möbius inversion for multivariant expressions

Before establishing bounds for comma-free dictionaries in multiple dimensions, we must establish Möbius inversion for multivariant expressions. Note that summing over multiple variables in the Möbius inversion formula

Lemma 6. $\sum_{d_i|k_i} f(d_1, d_2, \ldots, d_q) = g(k_1, k_2, \ldots, k_q)$ *is equivalent to*

$$f(k_1, k_2, \ldots, k_q) = \sum_{d_i|k_i} \left[\left(\prod_{i=1}^{q} \mu(k_i/d_i) \right) g(d_1, d_2, \ldots, d_k) \right]$$

Now that we have this formulation, we can proceed to our general bound for comma-free codes in multiple dimensions.

Theorem 8.

$$Q(k_1, k_2, \ldots, k_q, n) \leq \frac{\sum_{d_i|k_i} \left[\left(\prod_{i=1}^{q} \mu(k_i/d_i) \right) \left(\prod_{i=1}^{q} d_i \right) \right]}{\prod k_i}$$

Proof. We define a word with subperiod of size $d_1 \times d_2 \times \cdots \times d_q$ as a word formed by repeating a word of size $d_1 \times d_2 \times \cdots \times d_q$ to form a word of size $k_1 \times k_2 \times \cdots \times k_q$. We note that a word must have a subperiod of size $k_1 \times k_2 \times \cdots \times k_q$ to be in a comma-free dictionary. Otherwise, placing the word next to 2^q copies of itself yields the original word as an overlap.

Let $f(d_1, d_2, \ldots, d_q)$ be the number of words with subperiod of size $d_1 \times d_2 \times \cdots \times d_q$. All words of size $k_1 \times k_2 \times \cdots \times k_q$ must have some subperiod of size $d_1 \times d_2 \times \cdots \times d_q$ where $d_i | k_i$ for all i. The total number of words of size $k_1 \times k_2 \times \cdots \times k_q$ is $\prod_{i=1}^{q} k_i$. Thus,

$$\sum_{d_i | k_i} f(d_1, d_2, \ldots, d_q) = \prod_{i=1}^{q} k_i.$$

Using our formula for Möbius inversion for multivariant functions,

$$f(k_1, k_2, \ldots, k_q) = \sum_{d_i | k_i} \left[\left(\prod_{i=1}^{q} \mu(k_i/d_i) \right) \left(\prod_{i=1}^{q} d_i \right) \right].$$

Furthermore, we create equivalence classes of words which are equivalent under one or more cyclic shifts along any dimension. No two equivalent words can be in a comma-free dictionary, as repeating one word yields all equivalent words as an overlap. There are $\prod_{i=1}^{q} k_i$ words in each equivalence class. Thus we can divide by $\prod_{i=1}^{q} k_i$ to yield the maximum number of such words that can inhabit a comma-free dictionary. This gives us our desired result. □

4.1 Possible additional bounds

The bounds determined for q dimensions are not always tight. Indeed, there are several other cases which can be eliminated, though they are more difficult to classify. Specifically, it is possible to eliminate all words which are fixed under some nonidentity cyclic shift over q dimensions. This includes but is not limited to cyclic shifts along a single dimension. Subperiods can take place over multiple dimensions. For instance, in two dimensions, cyclic shifts of a repeated block of letters can yield a subperiod in two dimensions, as in the following example: $\begin{matrix} a_1 & a_2 & a_3 & a_4 \\ a_3 & a_4 & a_1 & a_2 \end{matrix}$. Since such matrices cannot be comma-free, they can improve existing bounds; however, their properties are inconsistent. This makes a tight bound difficult. For this reason, we have not utilized this observation to improve our bounds.

4.2 Self-reflectivity in multiple dimensions: implications and applications

Now we combine two original concepts in this paper: self-reflective comma-freeness and comma-free codes in multiple dimensions. We expand our definition of words in multiple dimensions to include arrays on a multidimensional lattice of size $k_1 \times k_2 \times \cdots \times k_q$ with orientation along any dimensional axis. We define a Multiorientational Comma-Free

Dictionary by requiring that if a multidimensional word ω is in our dictionary, so too must be all dimensional orientations of ω. Since there are q dimensions, there are thus 2^q words which are all possible orientations of any word. Since standard self-reflective comma-free codes are in one dimension, there are two orientations of any word: forward and backward. In other words, for each word in a multiorientational dictionary, its reflection must also be in that dictionary. Thus self-reflective comma-free dictionaries are the special case of multiorientational comma-free dictionaries for one dimension.

The implications of multiorientational comma-free dictionaries are staggering. By utilizing a single dimension for a standard word and filling the rest of a multidimensional word with a uniform extra character, it is possible to create a variable-size comma-free dictionary, as the size of a word in each dimension can contain as many as q different lengths in q directions. Variable-size comma-free dictionaries have even more surprising applications. Variable-size comma-free dictionaries have direct implications to the NP-complete Post Correspondence problem. If all words in the Post Correspondence problem were members of some variable-size comma-free dictionary, the problem would have no solutions. As this has implications to an undecidable decision problem, variable-size comma-free dictionaries have enormous implications in theoretical math. Comma-free codes in multiple dimensions also display potential for future coding and cryptographic techniques.

5 Conclusion

This work addresses the new problem of self-reflective comma-free codes. These codes address the critical problem of efficient use of stamp printing by a receiver. This work attempts to gain bounds on the size of self-reflective comma-free dictionaries given variable word-length and alphabet size. This work also discusses the new problem of self-swappable comma-free codes and the generalization to comma-free codes in multiple dimensions.

We achieve tight bounds for specific word-length and variable alphabet length, as well as general bounds for general word-length. The results are limited in scope to constructions under which a reflection is equivalent to a cyclic shift. We proceed to address other classes of comma-free codes including self-swappable codes and comma-free codes over q dimensions. We prove general bounds for these classes, but they contain many open problems. Future extensions of this project could include attempts at tight bounds for general word-length as well as efficient methods of construction for self-reflective comma-free codes. Improved bounds on comma-free codes in multiple dimensions should also be attempted.

Acknowledgements:

I would like to thank the Research Science Institute in coordination with the Massachusetts Institute of Technology for the opportunity to perform this research over Summer 2007. Additionally, I would like to thank Ms Amanda Epping Redlich, of MIT, for mentoring me and suggesting Comma-Free Codes as a field of study in coordination with Dr P. W. Shor and Dr David Jerison. Finally, I would like to thank Dr John Rickert for teaching me how to write a research paper and Harrison Chen and Allison Gilmore for helping to proofread the paper.

References

[1] K. L. Collins, P. W. Shor, and J. R. Stembridge. A Lower Bound for 0, 1, * Tournament Codes. *Discrete Math.* **63**, (1987) 15–19.

[2] F. H. C. Crick, J. S. Griffith, and L. E. Orgel. Codes Without Commas. *Proc. Nat. Acad. Sci.* **43** (1957), 416–421.

[3] W. L. Eastman. On the Construction of Comma-Free Codes. *IEEE Trans. Inform. Theory* **11** (1965), 263–266.

[4] S. W. Golomb, B. Gordon, and L. R. Welch. Comma-Free Codes. *Canad. J. of Math.* **10** (1958), 202–209.

[5] B. H. Jiggs. Recent Results in Comma-Free Codes. *Canad. J. Math.* **15** (1963), 178–187.

[6] V. I. Levenshtein. Combinatorial Problems Motivated by Comma-Free Codes. **J. Combin. Des. 12** (2004) 184–196.

[7] R. A. Scholtz. Maximal and Variable Word-Length Comma-Free Codes. *IEEE Trans. Inform. Theory* **15** (1969), 300–306.

Appendices

A Constructions for self-reflective comma-free codes

A.1 Self-reflective comma-free codes of word-length 4

We construct a self-reflective comma-free dictionary for $k = 4$ by including all words ABCD such that $A > B > C \leq D$ or $A \geq B < C < D$. This construction is self-reflective and comma-free. The size of the dictionary created by the construction over an n-letter alphabet is $\frac{n^4-2n^3-n^2+2n}{4}$. The bound from theorem 6 on the size of such a dictionary is $\frac{n^4-2n^3+n^2}{4}$. This differs from the size of the construction by $\binom{n}{2}$. It is interesting to note the size of the construction for $n = 4$. According to Levenshtein [6], the maximum size of a comma-free dictionary with $k = 4$ and $n = 4$ is 57. This is 3 less than bound (1) would predict. The size for a self-reflective comma-free dictionary under this construction for $n = 4$ is 30. This is 6 less than the bound from Theorem 6 would predict. It is possible that for each of the three words which could not fit into the 60-member dictionary, those words and their reflections must be eliminated from a self-reflective dictionary, yielding 30 words.

A.2 Self-reflective comma-free codes of odd word length

We construct a self-reflective comma-free dictionary for odd k by including all words $a_1 a_2 \cdots a_k$ such that

$$a_1 > a_2 > \cdots > a_t < a_{t+1} < \cdots < a_k$$

This construction is self-reflective and comma-free, but it is not a maximal construction for all k. Despite this, it is a convenient and consistent method of construction for self-reflective comma-free dictionaries.

B Other comma-free conjectures

B.1 Creating new dictionaries from existing dictionaries

One conjecture we addressed was the potential that for every comma-free dictionary $D = \omega_1, \omega_2, \ldots, \omega_x$, there exists another comma-free dictionary $D' = \phi(\omega_1), \phi(\omega_2), \ldots, \phi(\omega_x)$ created by taking a cyclic shift of each word in the dictionary. This is not necessarily true.

Without loss of generality, let $a_1 a_2 \cdots a_k \in D$ and $b_1 b_2 \cdots b_k \in D$. Now D' must contain $a_2 a_3 \cdots a_k a_1$ and $b_2 b_3 \cdots b_k b_1$, so D' cannot contain any of $a_3 a_4 \cdots a_k a_1 b_2$, $a_4 a_5 \cdots a_k a_1 b_2 b_3$, $\ldots$, $a_1 b_2 b_3 \cdots b_k$.

Thus D could not have contained any of $b_2 a_3 a_4 \cdots a_k a_1$, $b_3 a_4 a_5 \cdots a_k a_1 b_2$, and so on.

But it may be feasible to include some of these words in D, since they are not necessarily overlaps of the original two words. Thus if D is comma-free, D' is not necessarily comma-free.

B.2 Creating new dictionaries from existing dictionaries using half-shifts

While it is not possible to create new dictionaries from any cyclic shift of every word in the dictionary, it is possible to create new dictionaries using cyclic shifts of $\frac{k}{2}$ provided k is even. This is clear, because it is possible to consider each string of $\frac{k}{2}$ letters as a single letter over an alphabet of size $n^{k/2}$. Then a half-shift is equivalent to a reflection over that alphabet. If the original dictionary was comma-free, then this reflection will be comma-free, as the letters formed by words will remain comma-free.