On the Structure of Anagrammatic Encodings

LARRY E. STANFEL

Department of Industrial and Systems Engineering, University of Florida
Gainesville, Florida 32601

The anagrammatic, (never-synchronizing) property is investigated in terms of some structural aspects of exhaustive prefix encodings. In the case of alphabets having even numbers of characters, a condition necessary for the anagrammatic property is derived in terms of the number of second-longest words. Structure functions are used to derive a necessary condition for the property in terms of the sets of proper prefixes of an encoding and its reversal.

Anagrammatic encodings, shown to be the only class of never-synchronizing encodings (Gilbert and Moore, 1959), are of interest owing to this fact itself, as well as their apparent scarcity (Schutzenberger, 1956) relative to the set of all exhaustive, prefix codes. Not a great deal is known about their more specific structural properties, and this paper serves to identify several of these.

$E^m$ will denote the set of all strings of $m$ characters which may be formed from an alphabet $E$.

Juxtaposition will denote the operation of concatenation, and if $A = (a_1, \cdots, a_m), B = (b_1, \cdots, b_n)$,

$$AB = (a_1b_1, \cdots, a_nb_n, a_2b_1, \cdots, a_2b_n, \cdots, a_mb_n),$$

the $a_i$, $b_j$ being arbitrary strings over an alphabet $E$.

The superscript $R$ will be used to indicate the operation of reversal; for example, if $s = abc$, $s^R = cba$, for $a$, $b$, $c$ symbols in an alphabet. Further, if an encoding $W$ has code words, $w_1, w_2, \cdots, w_n$, then $W^R$ has code words $w_1^R, w_2^R, \cdots, w_n^R$.

**Lemma 1.** Let $W$ be an anagrammatic encoding over an alphabet $E$.

---

1 Research supported in part by Project THEMIS Contract AROD-DAH-CO4-68C-0002

Copyright © 1969 by Academic Press, Inc.
Suppose it has a set $C_i$ of code words of length $k_i$ for $i = 1, 2, \ldots, m$. Assume $k_1 < k_2 < \cdots < k_m$. Then there exists a set $S$ of strings such that

$$L = SE^{k_m-k_{m-1}}$$

and

$$L' = E^{k_m-k_{m-1}}S$$

where $L$ is the union of $C_{m-1}$ and the set of $k_{m-1}$—length strings eliminated from $W$ by the anagrammatic assumption and $L'$ is the union of $C_{m-1}$ and the set of $k_{m-r}$—length strings eliminated by the prefix property.

Proof. The set of $k_m$—length strings eliminated from $W$ by the prefix property is

$$F_1 = C_1E^{k_m-k_1} U C_2E^{k_m-k_2} U \cdots U C_{m-1}E^{k_m-k_{m-1}}.$$  

Similarly, the set of strings eliminated by the suffix (anagrammatic) property is

$$F_2 = E^{k_m-k_1}C_1 U E^{k_m-k_2}C_2 U \cdots U E^{k_m-k_{m-1}}C_{m-1}.$$  

We must have $F_1 = F_2$, or

$$C_1E^{k_m-k_1} U C_2E^{k_m-k_2} U \cdots U C_{m-1}E^{k_m-k_{m-1}} = E^{k_m-k_1}C_1 U E^{k_m-k_2}C_2 U \cdots U E^{k_m-k_{m-1}}C_{m-1}. \quad (1)$$

Equation (1) may be rewritten

$$(C_1E^{k_m-k_1-1} U \cdots U C_{m-1}) E^{k_m-k_{m-1}} = E^{k_m-k_{m-1}}(E^{k_m-k_1}C_1 U \cdots U C_{m-1}) \quad (2)$$

where the parenthesized sets are sets of strings of length $k_{m-1}$, and neither set consists of all possible strings of that length.

Equation (2) implies $k_{m-1} > k_m - k_{m-1}$, for equality could not hold for $k_{m-1} \leq k_m - k_{m-1}$. Of course, we know that if $k_i < k_m - k_{m-1}$ for any $i$, $1 \leq i \leq m - 1$, the encoding cannot be anagrammatic.)

At any rate (2) may be rewritten

$$E^{k_m-k_{m-1}}L = L'E^{k_m-k_{m-1}}. \quad (3)$$

The following observations are then possible. For each element of the set $L'$ there must exist a set of $E^{k_m-k_{m-1}}$ strings in $L$, each suffixed by a distinct $(k_m - k_{m-1})$—length string and each having its first

$$k_{m-1} - (k_m - k_{m-1}) = 2k_{m-1} - k_m$$
characters equal to the last \(2k_{m-1} - k_m\) characters of the \(L'\) element selected, where \(D\) is the cardinality of \(E\).

In similar fashion, for each element of \(L\) there must exist \(D^{k_m - k_{m-1}}\) elements of \(L'\), each prefixed by a distinct \((k_m - k_{m-1})\)-length string and each having its last \(2k_{m-1} - k_m\) characters equal to the first \(2k_{m-1} - k_m\) characters of the selected \(L\) element.

Thus, it must be possible to factor

\[
L' = E^{k_{m-1} - k_m} S_1,
\]

where \(S_1\) is a set of strings of length \(2k_{m-1} - k_m\).

Also, there must exist a set \(S_2\) having the same length strings as \(S_1\) and containing a number of strings equal to that of \(S_2\), and such that

\[
L = S_2 E^{k_{m-1} - k_m}.
\]

Substituting (4) and (5) into (3), we find

\[
E^{k_{m-1} - k_m} S_2 E^{k_{m-1} - k_m} = E^{k_{m-1} - k_m} S_1 E^{k_{m-1} - k_m}.
\]

Equation (6), in turn, implies that \(S_1 = S_2\).

Therefore, letting \(S = S_1 = S_2\), the lemma is proved.

With the help of Lemma 1, we may prove the following.

**Theorem 1.** If \(W\) is an anagrammatic encoding over an alphabet with an even number of characters with word lengths \(k_1 < k_2 < \cdots < k_{m-1} < k_m\), then there is an even number of words of length \(k_{m-1}\).

**Proof.** Let \(W\) be anagrammatic.

Since \(E\) has an even number of characters \(E^{k_{j-i} \cdot k_i}(k_j > k_i)\) is always even. Consequently, sets of the form \(C_iE^{k_{m-1} - k_i}C_i\) and \(E^{k_{m-1} - k_i}C_i\) in (2) have even numbers of elements. If then \(C_{m-1}\) has an odd number of elements, \(L\) and \(L'\) have an odd number of elements.

But then the set \(S_i\) in Lemma 1, cannot exist, for if it did, we would have the product of an even integer and another integer being equal to an odd integer, from (4) and (5). Since this is impossible, Theorem 1 is proved.

Consider next the structure functions as used in (Karp, 1961).

Let the alphabet \(E = (c_1, c_2, \cdots, c_e)\).
Let $n_{i,k}$ be the number of occurrences of $c_k$ in code word $w_i$, and let

$$f(w_i) = \prod_{j=1}^{k} s_j^{n_{i,j}}$$

$i = 1, \cdots, n$

with $W = (w_1, w_2, \cdots, w_n)$.

Define $F(W) = \sum_{m=1}^{n} f(w_m)$.

$F$ is called the multivariate structure function of $W$ in the variables $s_1, \cdots, s_k$.

Karp proved, then, the following theorems.

**Theorem A1.** $P(s_1, \cdots, s_k)$ is the multivariate structure function of an exhaustive prefix code if and only if

(a) $F(0, \cdots, 0) = 0$

(b) Every coefficient in the polynomial is nonnegative.

(c) $P(s_1, \cdots, s_k) - 1 = (s_1 + s_2 + \cdots + s_k - 1) Q(s_1, \cdots, s_k)$.

**Theorem A2.** If $P = F(W)$ with $W$ an exhaustive prefix encoding having distinct proper prefixes $p_1, p_2, \cdots p_n$ then

$$Q = \sum_{i=1}^{n} f(p_i),$$

$Q$ being the polynomial of Theorem A1.

Let $W$ be an anagrammatic encoding.

We have $F(W) = F(W^R)$, by definition of $F$.

From Theorem A1, (c)

$$F(W) = P(s_1, \cdots, s_k)$$

$$= (s_1 + \cdots + s_k - 1) Q(s_1, \cdots, s_k) + 1$$

$$= (s_1 + \cdots + s_k - 1) \cdot \sum_{i=1}^{n} f(p_i) + 1,$$

where $p_1, \cdots, p_n$ are the distinct proper prefixes of $W$.

Now, since $F(W) = F(W^R)$, it must be that

$$F(W^R) = (s_1 + \cdots + s_k - 1) \cdot \sum_{i=1}^{n} f(p_i) + 1;$$

but also

$$F(W^R) = (s_1 + \cdots + s_k - 1) \cdot \sum_{i=1}^{n} f(p_i') + 1.$$
by Theorem 2, where $p_1', \ldots, p_n'$ are the distinct proper prefixes of $W^R$.

It is clear, then, that not only do $W, W^R$ have an identical number of distinct proper prefixes, but also that those of $W^R, (p_1', p_2', \ldots, p_n')$ may be obtained from $(p_1, \ldots, p_n)$ by permuting the symbols of each of the latter set in some fashion or other. Stated in another manner, there exists a 1:1 correspondence between $(p_1, \ldots, p_n)$ and $(p_1', \ldots, p_n')$ such that if

$$p_i \text{ corresponds to } p'_j$$

then one is just a permutation of the other.

Thus we have proved.

**Theorem 2.** If $W$ is an anagrammatic encoding with distinct proper prefixes $(p_1, p_2, \ldots, p_n)$, then the distinct proper prefixes $(p_1', p_2', \ldots, p_n')$ of $W^R$ are in 1:1 correspondence with $(p_1, \ldots, p_n)$, and if $p_i$ corresponds to $p'_j$, then one is a permutation of the other.

As a referee has pointed out, stronger statements regarding the above relationship may be made.

First, consider a proper prefix $c_1c_2, \ldots, c_n$ in an anagrammatic code $W$, where the $c_j$ are symbols from the particular alphabet. Then there are words beginning $c_1c_2 \ldots c_n c_1$. If $c_1c_2 \ldots c_n c_1$ is a word, then $c_2 \ldots c_n c_1$ is a proper suffix of $W$. Otherwise, there are words beginning $c_1c_2 \ldots c_n c_1 c_2$. If that is a word, then $c_3 \ldots c_n c_1 c_2$ is a proper suffix of $W$. Continuing in this fashion, one finds that if $c_1c_2 \ldots c_n$ is a proper prefix of $W$, then one of

$$c_1 c_2 \ldots c_n$$
$$c_2 c_3 \ldots c_n c_1$$
$$c_3 \ldots c_n c_1 c_2$$
$$\vdots$$
$$c_{n-1} c_n c_1 \ldots c_{n-2}$$
$$c_n c_1 c_2 \ldots c_{n-1}$$

is a proper suffix of $W$; that is, for some $i$, we have $c_1 c_2 \ldots c_i c_{i+1} \ldots c_n$ a proper prefix of $W$ and $c_{i+1} \ldots c_n c_1 \ldots c_i$ a proper suffix of $W$. Since $W$ is anagrammatic, the same result may be obtained between proper prefixes of $W^R$ and proper suffixes of $W^R$, and, therefore, this relationship is one to one from proper prefixes of $W$ onto proper suffixes of $W$. 
Finally, let us show that this same correspondence exists between proper prefixes of $W$ and proper prefixes of $W^R$. Suppose $c_1c_2 \cdots c_n$ is a proper prefix in $W$ and that none of

$$c_1c_2 \cdots c_n$$
$$c_2 \cdots c_nc_1$$
$$c_3 \cdots c_nc_1c_2$$
$$\vdots$$
$$c_{n-1}c_nc_1 \cdots c_{n-2}$$
$$c_nc_1c_2 \cdots c_{n-1}$$

are proper prefixes of $W^R$. Then each must be prefixed by a word of $W^R$. If no string of the form $c_{i+1} \cdots c_nc_1 \cdots c_i$ is a proper prefix of $W^R$, then none of

$$c_{i+1} \cdots c_nc_1 \cdots c_i$$
$$c_{i+2} \cdots c_nc_1 \cdots c_ic_{i+1}$$
$$\vdots$$
$$c_{n}c_1 \cdots c_ic_{i+1} \cdots c_{n-1}$$
$$\vdots$$
$$c_i \cd_{i+1} \cdots c_nc_1 \cdots c_{i-1}$$

can be proper suffixes of $W^R$, as previously shown, which implies that none of

$$c_i \cdots c_ic_{n} \cdots c_{i+1}$$
$$c_{i+1}c_i \cdots c_{i+1}c_{n} \cdots c_{i+2}$$
$$\vdots$$
$$c_{i-1} \cdots c_{i}c_{n} \cdots c_{i+1}c_i$$

are proper prefixes of $W$, so that each string of $U$ is prefixed by a word of $W$. But since $U = V^R$, each element of $V$ must be prefixed and suffixed by a word of $W^R$, and the only way in which such a set may be constructed, preserving both the prefix and suffix properties within that set is to select a set of words of constant length. We will denote such a
Now it is also the case that (i) there is no element \( w \) of \( W^R \) such that \( w = \alpha \beta \) with \( \alpha \neq \phi, \beta \neq \phi \), and \( z \in Z \); for if say \( w = \alpha_1 \alpha_2 \cdots \alpha_k \cdot c_1 c_2 \cdots c_j \beta_1 \beta_2 \cdots \beta_r \), there must also be words of \( W^R \) beginning \( \alpha_1 \alpha_2 \cdots \alpha_k c_1 c_2 \cdots c_j c_{j+1} \).

This last string cannot be a word, since then \( c_2 \cdots c_{j+1}, \) a word, would suffix it. But then, there must be words beginning \( \alpha_1 \cdots \alpha_k \cdot c_1 c_2 \cdots c_{j+1} c_{j+2}, \) etc, and the suffix property cannot be preserved, since the string must eventually terminate, ending with an element of \( Z \).

It is clear, in addition, that (ii) no element \( z \) of \( Z \) may be of the form \( z = \alpha w \beta, \alpha \neq \phi, \beta \neq \phi, w \in W^R \). Otherwise, the prefix property is missing.

One way to observe that such an encoding cannot be anagrammatic is to attempt to construct it. First, there are no words of length less than \( j - 1 \), by (i), since \( W^R \) is assumed anagrammatic.

Construct, then, an exhaustive, prefix encoding \( W_0 \) over the given alphabet \( E \) such that every possible \( (j - 1) \) length string is a code word except the \( (j - 1) \) length proper prefixes of \( Z \), which are proper prefixes of \( j \) length words in \( W_0 \); that is, \( W_0 \) also has words \( c_1 c_2 \cdots c_j E, c_2 \cdots c_j c_k E, \ldots, c_j c_1 \cdots c_{j-1} E \).

Beginning with \( W_0 \), then, construct \( W_1, W_2, \cdots \), where \( W_{i+1} \) is obtained from \( W_i \) by selecting one word \( w_i^* \in W_i \) and replacing this in \( W_{i+1} \) by a set of words \( w_i^* E \).

By selecting the appropriate word \( w_i^* \) at each iteration, for some finite \( k \in \{0, 1, 2, \cdots \} \), we realize \( W_k = W^R \); but the elements of \( Z \) must be present in each of \( W_0, W_1, \cdots, W_{k-1}, W_k \), and because of this, and the fact that if a word of \( Z \) suffixes another word in \( W_i \) there will be a word of \( Z \) suffixing another in every subsequent encoding, every encoding in the sequence \( \{W_i\} \) has at least one word suffixing another, regardless of the choice of the \( w_i^* \), unless \( W^R \) is a uniform code in which case the prefix correspondence is clearly present. The only other exception is the case where the elements of \( Z \) are identical, in which case the
given prefix has the form \( cc \cdots c \) for some \( c \in E \), and for which the desired correspondence is also apparent.

The contradiction is thus established. Summarizing, we have

**Theorem 3.** If \( W \) is an anagrammatic encoding, then there is a 1:1 correspondence from proper prefixes of \( W \) onto proper suffixes of \( W \), such that if \( c_1c_2\cdots c_n \) is a proper prefix of \( W \), then for some \( i \), \( c_{i+1}\cdots c_n c_1\cdots c_i \) is a proper suffix of \( W \). Further, the same correspondence holds from proper prefixes of \( W \) onto proper prefixes of \( W^R \).

**Received:** March 10, 1967; revised: December 12, 1968

**References**

