

https://dspace.org/

Counting Distinct Adjacent rtuples in Words

Aubrey Blecher; Arnold Knopfmacher

CHARLES BABBAGE RES CTR
https://hdl.handle.net/10539/40918
SCOPUS

Downloaded from a DSpace repository. https://dspace.org/

Ars Combinatoria, 158: 93–106
DOI:10.61091/ars158-10
http://www.combinatorialpress.com/ars
Received 13 February 2024, Accepted 15 March 2024, Published 31 March 2024

Combinatorial Press

Article

Counting Distinct Adjacent r-tuples in Words

Aubrey Blecher[1,*], **Arnold Knopfmacher**[1]

[1] The John Knopfmacher Centre for Applicable Analysis and Number Theory, School of Mathematics, University of the Witwatersrand, Private Bag 3, Wits 2050, South Africa

* **Correspondence:** aubrey.blecher@wits.ac.za

Abstract: For $r = 1, 2, \ldots, 6$, we obtain generating functions $F_k^{(r)}(y)$ for words over the alphabet $[k]$, where y tracks the number of parts and $[y^n]$ is the total number of distinct adjacent r-tuples in words with n parts. In order to develop these generating functions for $1 \leq r \leq 3$, we make use of intuitive decompositions but for larger values of r, we switch to the cluster analysis method for decorated texts that was introduced by Bassino et al. Finally, we account for the coefficients of these generating functions in terms of Stirling set numbers. This is done by putting forward the full triangle of coefficients for all the sub-cases where $r = 5$ and 6. This latter is shown to depend on both periodicity and number of letters used in the r-tuples.

Keywords: Generating function, Set partitions of n into j blocks, Restricted growth functions, Stirling numbers of the second kind
2020 Mathematics Subject Classification: 05A18, 05A15

1. Introduction

A word of length n over an alphabet $[k] = \{1, 2, \ldots, k\}$ is an element of $[k]^n$. So such a word w is a sequence of n parts, where each part is in $[k]$. For $1 \leq r \leq n$, an adjacent r-tuple is a subsequence of r adjacent parts of w.

For example, the set of words over the alphabet [2] of length 3 is

$$\{111, 112, 121, 122, 211, 212, 221, 222\}.$$

In the above example, the 2-tuples are given by $\{11, 12, 22, 21\}$. The number of occurrences of each of these adjacent 2-tuples is as follows: 11 occurs in 3 of the words of length 3. 22 occurs also in 3 of the words of length 3. 12 occurs in 4 of these. 21 occurs also in 4. The word 112 contains two distinct 2-tuples, namely 11 and 12, whereas 111 contains one 2-tuple, namely 11. Altogether the set of words of length three contains a a total of $14 = 3 + 3 + 4 + 4$ distinct 2-tuples.

The main aim of this paper is as follows: In the case of $1 \leq r \leq 6$, we will develop a generating function $F_k^{(r)}(y)$ for words over the alphabet $[k]$; where y tracks the number of parts and the coefficient of y^n, i.e., $[y^n]F_k^{(r)}$, is the sum of the number of different words with n parts counted once for each distinct adjacent r-tuple it contains. The complexity of the problem increases significantly as r grows. A brief discussion of r-tuples for $r > 6$ is given in the final section.

Thus $[y^3]F_2^{(2)}(y) = 14$. Similarly $[y^3]F_3^{(2)}(y) = 51$ since each of 11, 22 and 33 occur in 5 words of length 3 and each of 12,13,21,23,31,32 occur in 6 words of length 3.

Analogous problems have been studied before by the current authors and others in several different contexts. So, for example, in geometrically distributed words, we have studied the number of distinct adjacent pairs, see [1]. And in integer partitions of n, we have studied the number of distinct r-tuples, see [2].

All these problems are a generalisation of the old initially studied problem, namely, the number of distinct parts in a given combinatorial structure. For a recent presentation of this problem in the case of integer partitions, see for example, [3]. Whereas the number of distinct letters in geometrically distributed words is studied in [4].

To build an insight into the problem we will start by developing the generating functions for adjacent pairs (in section 2) and adjacent triplets (in Section 3). For these, we use basic and intuitive decompositions. For 4-tuples in Section 4, we introduce the cluster analysis method in the paper by Bassino et al. [5]. The cluster analysis method is also applied to larger values of r in Sections 5 and 6. In Sections 7 and 8, we account for the coefficients of these generating functions in terms of Stirling set numbers. This is done by putting forward the full triangle of coefficients for all the sub-cases where $r = 5$ and 6. This latter is shown to depend on both periodicity of the subwords (r-tuples) being counted and number of letters used in these.

2. Generating Functions for Total Number of Distinct Adjacent Pairs in Words

We set out to construct the generating function $F_k^{(2)}(y)$ for words over the alphabet $[k]$; where y tracks the number of parts and $[y^n]$ is the total number of distinct two letter sequences in words with n parts.

In order to obtain an expression for $F_k^{(2)}(y)$, for $a \neq b$, we first let $f_{ab}(y)$ be the generating function for all words with y marking the number of parts, which contain the pair ab; $\bar{f}_{ab}(y)$ be the complementary function of f_{ab} (i.e., it tracks words which do not contain ab); $f_{aa}(y)$ be the generating function for all words with y marking number of parts, which contain the pair aa and $\bar{f}_{aa}(y)$ its complementary function.

To obtain the latter generating functions we set out a few symbolic decompositions.

Firstly, for words which avoid ab, we let S be a sequence of possibly empty non-$a's$ and we let c denote a single letter that is neither a nor b, which leads to the decomposition for such words as $(Sa^+c)^*Sa^*$, where * and + have the usual meaning for symbolic decompositions, namely they indicate respectively, possibly empty and non-empty sequences.

From this decomposition, we obtain its generating function

$$\bar{f}_{ab}(y) = \frac{1}{1 - \frac{1}{1-(k-1)y}\frac{y}{1-y}(k-2)y} \frac{1}{1-(k-1)y} \frac{1}{1-y} = \frac{1}{1-ky+y^2}. \tag{1}$$

Hence its complementary generating function is given by

$$f_{ab}(y) = \frac{1}{1-ky} - \bar{f}_{ab}(y) = \frac{1}{1-ky} - \frac{1}{1-ky+y^2}. \tag{2}$$

Secondly, the symbolic decomposition for words avoiding aa, where S is as before, $\bar{a}$ denotes a single letter that is not a and ϵ denotes the empty word, is $(Sa\bar{a})^*S(\epsilon + a)$. This yields the generating function

$$\bar{f}_{aa}(y) = \frac{1}{1 - \frac{1}{1-(k-1)y}y(k-1)y} \frac{1}{1-(k-1)y}(1+y) = \frac{1+y}{1+y-ky+y^2-ky^2}, \tag{3}$$

from which it follows that

$$f_{aa}(y) = \frac{1}{1-ky} - \bar{f}_{aa}(y) = \frac{1}{1-ky} - \frac{1+y}{1+y-ky+y^2-ky^2}. \tag{4}$$

Now we are in a position to find

$$F_k^{(2)}(y) = \sum_{a\neq b} f_{ab}(y) + \sum_{a=1}^{k} f_{aa}(y) = k(k-1)f_{ab}(y) + kf_{aa}(y). \tag{5}$$

The two sums above follow because the summands are independent of the particular terms ab or aa.

Substituting (2) and (4) into (5) completes the proof our following theorem;

Theorem 1. *The generating function for the number of distinct pairs in words of length n tracked by y, over an alphabet* $[k]$ *is given by*

$$F_k^{(2)}(y) = (k-1)k\left(\frac{1}{1-ky} - \frac{1}{1-ky+y^2}\right) + k\left(\frac{1}{1-ky} - \frac{y+1}{1+y-ky-ky^2+y^2}\right).$$

2.1. Generating Function for the Number of Distinct Letters in Words

A simple adaptation of the above argument yields

$$F_k^{(1)}(y) = k\left(\frac{1}{1-ky} - \frac{1}{1-(k-1)y}\right).$$

3. Generating Functions for Each Case of Distinct Adjacent Triplets in Words

Although the methods of the previous section are also applicable here, it is sometimes simpler to calculate the generating functions directly rather than first obtaining its complementary function. We will be guided by simplicity as to which method to choose. Also, by considering all words in reverse order it is clear that $f_{abb}(y)$ has the same generating function as $f_{aab}(y)$.

3.1. The Triplet abc

Here a, b, c represent distinct letters, $[k]$ is the alphabet for what follows in this section, $f_{abc}(y)$ is the generating function for all words which contain the subword abc, $\bar{n}(abc)$ is the set of all words which do not contain abc as a subword, $n(abc)$ is the set of all words which have abc as a subword and W is the set of all words.

So according to the first occurrence of the subword abc, which is shown below immediately preceding the W, we have the decomposition equation, $n(abc) = \bar{n}(abc)abcW$, with generating function equation $f_{abc}(y) = (\frac{1}{1-ky} - f_{abc}(y))y^3\frac{1}{1-ky}$ and solution

$$f_{abc}(y) = \frac{y^3}{(1-ky)(1-ky+y^3)}. \tag{6}$$

3.2. The Triplets aab and abb

Using the notation already introduced and the same method as in the previous case, we obtain the same generating function

$$f_{aab}(y) = \frac{y^3}{(1-ky)(1-ky+y^3)}. \tag{7}$$

By considering words in reverse order, it follows that $f_{aab}(y) = f_{abb}(y)$.

3.3. Bijection

Generating functions (6) and (7) are the same which implies that there is a bijection between words of a certain length which contain the pattern abc and such words which contain the pattern aab. Here is such a bijection: any word which contains both subwords is mapped to itself; and any word which

contains one but not the other has all parts preserved except that all instances of the particular subword (say *aab*) is replaced by the other (say *abc*). This mapping is a bijection between the two sets of words. So for example the image of *abcaab* is itself whilst *abcāabc* is mapped to *aabāaab*.

3.4. The Triplet aaa

Extending the notation from before, $f_{aaa}(y)$ is the generating function for all words which contain the subword *aaa*, $\bar{n}(aaa)$ is the set of all words which do not contain *aaa* as a subword, $n(aaa)$ is the set of all words which have *aaa* as a subword, $\bar{a}$ is any letter other than a and as before the alphabet is $[k]$.

According to the first occurrence of the subword *aaa*, such words may be decomposed as either $aaaW$ or $\bar{n}(aaa)\bar{a}aaaW$, which leads to the generating function equation

$$f_{aaa}(y) = \left(1 + (k-1)y\left(\frac{1}{1-ky} - f_{aaa}(y)\right)(k-1)y\right)y^3\frac{1}{1-ky},$$

and solution

$$f_{aaa}(y) = \frac{y^3(1-y)}{(1-ky)\left((1-ky)+(k-1)y^4\right)}. \tag{8}$$

3.5. The Triplet aba

Let d be any single letter and cd be any pair of letters. Using a similar recursive method to the case above we have the decomposition equation $n(aba) = (\epsilon; d; \bar{n}(aba)(cd - \{ab\}))\, abaW$ where $(p_1; p_2; p_3)$ uses the convention that the semi-colons separate different possibilities p_1, p_2 and p_3 for the bracketed term, and $(cd - \{ab\})$ denotes an arbitrary pair of letters other than ab, with corresponding generating function equation

$$f_{aba}(y) = \left(1 + ky + (k^2-1)y^2\bar{f}_{aba}(y)\right)y^3\frac{1}{1-ky}.$$

We recall that $\bar{f}_{aba} = \frac{1}{1-ky} - f_{aba}(y)$ which leads to the solution

$$f_{aba}(y) = \frac{y^3(1-y^2)}{(1-ky)(1-ky+k^2y^5-y^5)}. \tag{9}$$

3.6. Generating Functions for Total Number of Distinct Adjacent Triplets in Words

In a similar way to that in Section 2, we set out to construct the generating function $F_k^3(y)$ for words over the alphabet $[k]$; where y tracks the number of parts and $[y^n]F_k^3(y)$ is the total number of distinct three letter sequences in words with n parts. Recall also as shown above that $f_{aab}(y) = f_{abb}(y)$.

We have

$$\begin{aligned} F_k^{(3)}(y) &= \sum_{a=1}^{k} f_{aaa}(y) + \sum_{a\neq b\neq c\neq a} f_{abc}(y) + 2\sum_{a\neq b} f_{aab}(y) + \sum_{a\neq b} f_{aba}(y) \\ &= kf_{aaa}(y) + k(k-1)(k-2)f_{abc}(y) + 2k(k-1)f_{aab}(y) + k(k-1)f_{aba}(y). \end{aligned} \tag{10}$$

Substitution of the particular r-tuple generating functions into the above equation results in our next theorem;

Theorem 2. *The generating function for the number of distinct triplets in words of length n tracked by y, over an alphabet* $[k]$ *is given by*

$$\begin{aligned} F_k^{(3)}(y) = &-\frac{ky^3\left(k^4y^2\left(y^4+y^3+2y^2+y+1\right)-ky^2\left(y^2+y-1\right)-y\left(y^3+1\right)\right)}{(ky-1)\left(ky-y^3-1\right)\left((k-1)y^3+(k-1)y^2+(k-1)y-1\right)\left(-k\left(y^3+y\right)+y^3+y^2+1\right)} \\ &-\frac{ky^3(k^2(2y^6+4y^5+4y^4+4y^3+2y^2+y+1)-k^3y(3y^5+4y^4+5y^3+4y^2+2y+2))}{(ky-1)\left(ky-y^3-1\right)\left((k-1)y^3+(k-1)y^2+(k-1)y-1\right)\left(-k\left(y^3+y\right)+y^3+y^2+1\right)}. \end{aligned}$$

Hence forward we will not produce the increasingly complicated explicit equations in y but for the remaining $F_k^{(r)}(y)$ generating functions, for larger r, we will stop at the form corresponding to that of (10).

The simplest of these formulas is $F_2^{(3)}(y)$ which gives

$$F_2^{(3)}(y) = -\frac{2y^3\left(5y^4 - 2y^3 + 10y^2 - 13y + 4\right)}{(y-1)(2y-1)\left(y^2+y-1\right)\left(y^3-y^2+2y-1\right)\left(y^3+y^2+y-1\right)}.$$

4. Generating Functions for Total Number of Distinct Four Part Subwords

The methods of the previous section are also applicable here. As an alternative to using decompositions, we introduce the cluster analysis method in the paper by Bassino et al. [5]. Two features in that paper of which we make extensive use is the counting of clusters of one word patterns and also the automated method applied to the reduced case of word counting by inclusion-exclusion.

At this stage our analysis is facilitated by the following definition: Given an arbitrary word $a_1a_2\ldots a_n$, if $a_1a_2\ldots a_s$ where $1 \le s \le n$ is the *smallest* subword such that $a_1a_2\ldots a_n = (a_1a_2\ldots a_s)^+a_1a_2\ldots a_j$ where $0 \le j < s$, then $a_1a_2\ldots a_n$ is said to be of *minimal period* $p(s)$. So for example 111 is $p(1)$, 121 is $p(2)$, 123 is $p(3)$, $12121 = (12)^21$ is $p(2)$ and $121211 = (12121)^11$ is of minimal period $p(5)$.

Any restricted growth function of minimal period $p(j)$ is also said to have period m for any multiple m of $p(j)$.

It is essential in these examples that the patterns are represented as restricted growth functions (described below), because later in our analysis, we will replace, say the pattern 12 by pj, where p and j are any of the $k(k-1)$ pairs of two distinct elements in the alphabet $[k]$, and the unique count of all these different possibilities requires the base case (a restricted growth function) to be well defined.

To explain the notion of a restricted growth function, we need the following concepts.

Firstly, a partition Π of set $[n] = \{1, 2, \ldots, n\}$ of size j (i.e., having j blocks) is a collection $B_1, B_2, \ldots, B_j$ that satisfies: $\emptyset \ne B_i \subseteq [n]$, $B_{i_1} \cap B_{i_2} \ne \emptyset$ for $i_1 \ne i_2$ and $\cup_{i=1}^{j} B_i = [n]$. The elements B_i are called blocks and we list these blocks in increasing order of their minimal elements, that is, $\min B_1 < \min B_2 < \ldots < \min B_j$. We denote the set of all partitions of $[n]$ with exactly j blocks by $P_{n,j}$ and $|P_{n,j}| := S(n, j)$, which are known as the Stirling set numbers or Stirling numbers of the second kind. A partition Π can be written as $\pi_1, \pi_2 \ldots \pi_j$, where $\pi_i \in B_i$ for all i. The latter form is called a restricted growth function, see [6]. The number of partitions of an n element set into j blocks is in bijection with restricted growth functions of length n using j different letters. For example, $\Pi = \{\{13\}, \{2\}, \{4\}\}$ is a partition of $[4]$ and the restricted growth function form is $\Pi = 1213$. It may already be clear that a restricted growth function has a 1 in position 1 and any entry is a positive integer at most one larger than the maximum of the entries in the positions to its left.

So consider the case of an r-tuple with period s where $1 \le s \le r$ and assume for the moment that such r-tuples all have the same generating function independent of which subword pattern they are counting. Let $f_{r,s}$ and $\bar{f}_{r,s}$ respectively be the generating functions for all words of any length which contain or avoid such clusters (i.e., r-tuples). Directly from page 39:10 of [5], we have that the generating function $T(z, t)$ for all words avoiding the cluster of r-tuples for which ξ is the generating function is given by

$$T(z, t) = \frac{1}{1 - A(z) - \xi(z, t)}, \tag{11}$$

where z marks the length of the word and t marks decorated texts as per Bassino [5], ξ is the generating function for the set of r-tuples and $A(z)$ is the generating function for the alphabet.

In our case $y = z$ and $A(y) = ky$ because $[k]$ is the alphabet. For any choice of the subword pattern that we wish to avoid, we have already assumed above that they all have the same generating function, here called $\xi(y, t)$. For any such representative r-tuple of period s, we define m to be the integer that

satisfies $ms \geq r > (m-1)s$, and again using page 39:10 of [5], the generating function for the single word decorated texts is given by

$$\xi(y,t) = \frac{y^r}{1 - t(y^s + y^{2s} + \ldots + y^{(m-1)s})}, \tag{12}$$

because the possible number of repeats of the periodic pattern is tracked by t. This is equivalent in [5] to tracking the number of decorated texts. Above, the number of repeats is any number from the set $\{1, 2, \ldots, m-1\}$ with generating function $y^s + y^{2s} + \ldots + y^{(m-1)s}$.

We want to count words avoiding a particular r-tuple which as explained in [5] we obtain by setting $t = -1$.

Hence the generating function is

$$\bar{f}_{r,s} = \frac{1}{1 - ky + \frac{y^r}{1+y^s+y^{2s}+\ldots+y^{(m-1)s}}}, \tag{13}$$

where $ms \geq r > (m-1)s$.

Example 1.

$$\bar{f}_{6,1} = \frac{1}{1 - ky + \frac{y^6}{1+y+\ldots+y^5}}, \bar{f}_{6,2} = \frac{1}{1 - ky + \frac{y^6}{1+y^2+y^4}}, \bar{f}_{8,3} = \frac{1}{1 - ky + \frac{y^8}{1+y^3+y^6}}.$$

So, e.g. $\bar{f}_{6,2}$ is the generating function for all words of any length which do not contain a 6-tuple of period 2. In the case $k = 3$ the complementary series begins $f_{6,2} = y^6 + \mathbf{6y^7} + 26y^8 + 102y^9 + 378y^{10} + \cdots$. The coefficient of the term in bold counts the 6 words of length 7 that contain $(12)^3$ namely $a(12)^3$ or $(12)^3a$ where a is any of 1,2,3.

We turn now to the case $r = 4$.

In the following subsections, a, b, c, d represent distinct letters, and any assumptions and notation is an obvious extension from the subsection on the triplet abc.

4.1. The Subwords aaab, aabb, aabc, abac, abbb, abbc, abcb, abcc, abcd

All these patterns are of minimal period $p(4)$ and have the same generating function obtained from (11), namely

$$\bar{f}_{4,4} = \frac{1}{1 - ky + y^4}. \tag{14}$$

We prove that each of these subwords has the same generating function by means of an adaptation of the bijection we used for the three letter patterns:

Define a bijection between words of a certain length which contain the pattern $abcd$ and such words which contain any other of the patterns w as follows: any word which contains both subwords is mapped to itself; and any word which contains one but not the other has all parts preserved except that all instances of $abcd$ are replaced by w and vice versa. This mapping is a bijection between the two sets of words.

4.2. The Subword aaba, abaa, abba, abca

All these patterns are $p(3)$ and from (13) (and an adaptation of the bijection above) have the same generating function; namely

$$\bar{f}_{4,3} = \frac{1}{1 - ky + \frac{y^4}{1+y^3}}. \tag{15}$$

Alternatively, we built the set of clusters (repeats of the periodic pattern) from the decomposition $abca(bca)^*$ where the initial r-tuple is tracked by y^4 and the sequential follow up $(bca)^*$ gives rise to $\frac{1}{1+y^3}$.

4.3. The Subword abab

This is the only p(2) pattern and again, we focus on finding $\bar{f}_{4,2} = \bar{f}_{abab}$.

$$\bar{f}_{4,2} = \frac{1}{1 - ky + \frac{y^4}{1+y^2}}. \tag{16}$$

Above, we built the set of periodic pattern repeats from the decomposition $abab(ab)^*$ and its generating function is adjusted accordingly.

4.4. The Subword aaaa

This is the only $p(1)$ pattern with solution

$$\bar{f}_{4,1} = \frac{1}{1 - ky + \frac{y^4}{1+y+y^2+y^3}}. \tag{17}$$

This time we have built the set of periodic pattern repeats from the decomposition $aaaa(a)^*$ and its generating function is adjusted accordingly.

4.5. The Connection between the Periodic Cases and Restricted Growth Functions

Note that for the purposes of obtaining the F generating function we need a unique way to count each subword. We achieve this by insisting that the subword be a restricted growth function. Any variation from this is uniquely counted by multiplying this by the appropriate falling factor coefficient which then uniquely counts all the variations from the subword actually being a restricted growth function. We explain this connection by means of an example. Consider the $p(3)$ subword *abba*. This is in bijection with the restricted growth function 1221, where the letter a is mapped to 1 and b is mapped to 2. A non restricted growth function like 1331 is taken care of by the $k(k-1)$ factor that precedes the 1221 generating function.

4.6. Generating Functions for Total Number of Distinct Four Part Subwords

Here we set out to construct the generating function $F_k^{(4)}(y)$ for words over the alphabet $[k]$; where y tracks the number of parts and $[y^n]F_k^{(4)}(y)$ is the total number of distinct four letter sequences in words with n parts. The example *abba* given in the previous section could be replaced by *baab*. This is another subword that needs to be counted in our overall generating function $F_k^{(4)}(y)$. As already shown, any word of type $p(s)$ has the same generating function. If, however, the word has l distinct letters, then it contributes $(k)_l$ (falling factorial) cases to the overall generating function (ie this is reflected as a coefficient in the generating function). Summing over all these terms we obtain our next theorem for the overall generating function:

Theorem 3. *The generating function for the number of distinct* 4*-tuples in words of length n tracked by y, over an alphabet* $[k]$ *is given by*

$$\begin{aligned} F_k^{(4)}(y) = {} & kf_{4,1} + k(k-1)f_{4,2} + (3k(k-1) + k(k-1)(k-2))f_{4,3} \\ & + (3k(k-1) + 5k(k-1)(k-2) + k(k-1)(k-2)(k-3))f_{4,4}. \end{aligned}$$

For the coefficient of $f_{4,3}$ above, there are two falling factorials each with its own numerical coefficient. The number of permutations possible when two different letters are chosen from the alphabet $[k]$ is $k(k-1)$. The coefficient 3 for this falling factorial is the number of restricted growth functions of length 4 with two distinct letters and ending in 1, namely 1221, 1211 and 1121. This number is the same as the Stirling set-number $S(3,2) = 3$.

For the coefficient of $f_{4,4}$ above, there are three falling factorials each with its own numerical coefficient. The number of permutations possible when two different letters are chosen from the alphabet $[k]$ is again $k(k-1)$. The coefficient 3 for this falling factorial is the number of restricted growth functions of length 4 and period 4 using two distinct letters. This latter condition requires the last letter not to be a 1 (so as to exclude period 3), and also excludes the period 2 case 1212. This means the cases counted are 1222, 1122 and 1112. This is again 3 cases.

The number of permutations possible when three different letters are chosen from the alphabet $[k]$ is $k(k-1)(k-2)$. The coefficient 5 for this falling factorial is the number of restricted growth functions of length 4 and period 4, this time using three distinct letters. This latter condition requires the last letter not to be a 1 (so as to exclude period 3), and also excludes the period 2 case 1212. This means the cases counted are 1123, 1213, 1223, 1232 and 1233. This yields 5 cases.

The number of permutations using 4 letters is $(k)_4$. There is only one restricted growth function of length 4, using 4 distinct letters. Hence the coefficient is 1.

5. Generating Functions for each of the Distinct Five Part Subwords

5.1. All p(5)subwords

Here a, b, c, d, e represent distinct letters. There are a total of 34 p(5) subwords, 1 of which uses 5 distinct letters; 9 which use 4 distinct letters; 18 which use 3 distinct letters and 6 which use 2 distinct letters. In Section 7, we will explain how to compute the coefficients $6, 18, 9, 1$ that arise in the previous sentence as well as the coefficients arising in the subsections below.

There is a bijection between words of a certain length which contain any particular one of these patterns and such words which contain any other one of the patterns. It is the same bijection as that defined in the four part subword subsection. Consequently all of these have the same generating function as $f_{5,5}(y)$. By generalising the argument given for $f_{4,4}(y)$, we obtain the generating function

$$\bar{f}_{5,5}(y) = \frac{1}{1 - ky + y^5}. \tag{18}$$

5.2. All p(4)subwords

Here there is a total of 12 $p(4)$ subwords, 1 of which has 4 distinct letters; 6 have 3 distinct letters; and 5 have 2 distinct letters. Using the same bijection as above, these all have the same generating function. Using the decomposition for one of these subwords, namely *abcda*, we obtain $\bar{n}(abcda) = \bar{n}(abcd) + \bar{n}(abcd)abcd\,(\epsilon + a^*\bar{n}(abcda))$. In terms of generating functions, we obtain

$$\bar{f}_{5,4}(y) = \bar{f}_{4,4}(y)\left(1 + y^4 + y^4(k-1)y\bar{f}_{5,4}(y)\right).$$

Based on this equation, the generating function for these subwords is

$$\bar{f}_{5,4}(y) = \frac{1 + y^4}{1 - ky + y^4 + y^5 - ky^5}. \tag{19}$$

5.3. All p(3)subwords

There are 4 such subwords: 3 have 2 distinct letters (these are *aabaa*, *abaab* and *abbab*) and 1 uses 3 distinct letters. Unlike all previous analysis, the three cases of two distinct letters have two different generating functions. We show the derivation for each of these here: For the case of three distinct letters, $\bar{n}(abcab)$ has decomposition

$$\bar{n}(abcab) = \bar{n}(ab) + \bar{n}(ab)ab\bar{n}(ab) + \bar{n}(ab)\,(ab\bar{n}(ab - c))^+\, ab\bar{n}(ab), \tag{20}$$

which we explain below.

The first term is the case of no ab's; the second has exactly one ab and the last has two or more ab's between which we have excluded a single c.

In terms of generating functions, we obtain

$$\bar{f}_{abcab}(y) = \frac{1+y^3}{1+y^3+y^5-k\,(y^4+y)}. \tag{21}$$

As already stated the three cases of two distinct letters have two different generating functions. We show the derivation for each of these here;

Firstly, $\bar{n}(aabaa) = \bar{n}(aaba) + \bar{n}(aaba)aaba(\epsilon + \bar{a}\bar{n}(aabaa))$, where $aaba$ in the second term on the right hand side shows the first occurrence of $aaba$ which cannot be followed immediately by a. Using the previous generating function for avoidance of $aaba$, we obtain

$$\bar{f}_{aabaa}(y) = \frac{1+y^3+y^4}{1+y^3+y^4+y^5-k\,(y+y^4+y^5)}. \tag{22}$$

Secondly, by adapting decomposition (20),

$$\bar{n}(abbab) = \bar{n}(ab) + \bar{n}(ab)ab\bar{n}(ab) + \bar{n}(ab)\,(ab(\bar{n}(ab) - b))^{+}\, ab\bar{n}(ab),$$

which has the same generating function solution as Eq. (21), namely

$$\bar{f}_{abbab}(y) = \frac{1+y^3}{1+y^3+y^5-k\,(y^4+y)} = \bar{f}_{abcab}(y). \tag{23}$$

Remark 1. *For the first time in this paper, Eqs (22) and (23) are different generating functions for tuples which have the same period. Both these generating functions are for subwords which use precisely 2 distinct letters. As an alternative way of deriving Eqs (22) and (23), we have the decompositions for periodic repeats (i.e., decorated texts in the language of [5]) for aabaa as aabaa$(baa+abaa)^*$ as well as for periodic repeats of abbab which is abbab$(bab)^*$.*

5.4. All p(2)subwords

There is only one such subword, with generating function

$$\bar{f}_{5,2}(y) = \frac{1+y^2+y^4}{1+y^2+y^4+y^5-k(y+y^3+y^5)}, \tag{24}$$

as derived from Eq. (13).

5.5. All p(1)subwords

There is only one such subword, with generating function

$$\bar{f}_{5,1}(y) = \frac{1}{1-ky+\frac{y^5}{1+y+y^2+y^3+y^4}}, \tag{25}$$

also as obtained from Eq. (13).

5.6. Generating Functions for Total Number of Distinct Five Part Subwords

We reiterate the procedure applied in the case of the four part subwords and this time obtain

Theorem 4. *The generating function for the number of distinct* 5*-tuples in words of length n tracked by y, over an alphabet* $[k]$ *is given by*

$$F_k^{(5)}(y) = kf_{5,1}(y) + (k)_2 f_{5,2}(y) + k^2(k-1)f_{abbab}(y) + (k)_2 f_{aabaa}(y)$$

$$+\Big((k)_4+6(k)_3+5(k)_2\Big)f_{5,4}(y)+\Big((k)_5+9(k)_4+18(k)_3+6(k)_2\Big)f_{5,5}(y),$$

where

$$f_{5,1}(y)=\frac{1}{1-ky}-\frac{1}{1-ky+\frac{y^5}{1+y+y^2+y^3+y^4}};$$

$$f_{5,2}(y)=\frac{1}{1-ky}-\frac{1+y^2+y^4}{1+y^2+y^4+y^5-k\left(y+y^3+y^5\right)};$$

$$f_{abbab}=f_{abcab}=\frac{1}{1-ky}-\frac{1+y^3}{1+y^3+y^5-k\left(y+y^4\right)};$$

$$f_{aabaa}=\frac{1}{1-ky}-\frac{1+y^3+y^4}{1+y^3+y^4-k\left(y+y^4+y^5\right)};$$

$$f_{5,4}(y)=\frac{1}{1-ky}-\frac{1+y^4}{1-ky+y^4+y^5-ky^5}$$

and

$$f_{5,5}(y)=\frac{1}{1-ky}-\frac{1}{1-ky+y^5}.$$

The expression simplifies to

$$F_k^{(5)}(y)=kf_{5,1}(y)+\left(k^2-k\right)f_{5,2}(y)+\left(k^3-k^2\right)f_{abbab}+\left(k^2-k\right)f_{aabaa}+\left(k^4-2k^2+k\right)f_{5,4}(y)$$
$$+\left(k^5-k^4-k^3+k^2\right)f_{5,5}(y). \tag{26}$$

6. Generating Functions for Total Number of Distinct Six Part Subwords

To begin, for period 3, there are three cases of two distinct letters, which have two different generating functions. We show the derivation for each of these here;

Firstly, subwords *aabaab* and *abbabb*, which have the same generating functions. Using [5], the decompositions for the decorated text avoiding these subwords is $aabaab(aab)^*$ and $abbabb(abb)^*$.

The generating functions obtained are:

$$\bar{f}_{aabaab}(y)=\bar{f}_{abbabb}(y)=\frac{1+y^3}{(1+y^3)(1-ky)+y^6}. \tag{27}$$

For the case of words avoiding *abaaba* we also use the method in [5] where here the decorated text avoiding decomposition is $abaaba(aba+baaba)^*$. The generating function obtained is

$$\bar{f}_{abaaba}(y)=\frac{1+y^3+y^5}{(1+y^3+y^5)(1-ky)+y^6}. \tag{28}$$

Similarly, for period 4 we have subwords *aaabaa* and *abaaab* which have different decompositions for avoidance. Using [5], the decorated text decomposition for avoiding *aaabaa* is $aaabaa(abaa+aabaa)^*$ which leads to generating function

$$\bar{f}_{aaabaa}(y)=\frac{1+y^4+y^5}{1+y^4+y^5+y^6-k\left(y+y^5+y^6\right)}. \tag{29}$$

For avoiding *abaaab*, the decorated text decomposition is $abaaab(aaab)^*$ which leads to generating function

$$\bar{f}_{abaaab}(y)=\frac{1+y^4}{1+y^4+y^6-k\left(y+y^5\right)}. \tag{30}$$

In the same vein, for period 4 with 3 letters we have, subwords *aabcaa* and *abccab* which have different decompositions for avoidance. Firstly the decorated text decomposition for *aabcaa* is $aabcaa(bcaa + abcaa)^*$ leading to generating function:

$$\bar{f}_{aabcaa}(y) = \frac{1 + y^4 + y^5}{1 + y^4 + y^5 + y^6 - k\left(y + y^5 + y^6\right)}, \tag{31}$$

which is the same formula as that for $\bar{f}_{aaabaa}$ given in Eq. (29). Whereas, for *abccab*, the avoidance case decorated text decomposition is $abccab(ccab)^*$ and hence we obtain

$$\bar{f}_{abccab}(y) = \frac{1 + y^4}{1 + y^4 + y^6 - k\left(y + y^5\right)}. \tag{32}$$

Finally, for period 4 with 4 letters, for the subword *abcdab*, we have by the same methods that the generating function is

$$\bar{f}_{abcdab}(y) = \frac{1 + y^4}{1 + y^4 + y^6 - k\left(y + y^5\right)}. \tag{33}$$

For the rest of this section we use methods already explained in the previous sections and therefore simply record the results.

The 6 cases for the gfs containing patterns of periods 1 up to 6 are
$\{f_{6,1}(y), f_{6,2}(y), f_{6,5}(y), f_{6,6}(y)\} = \left\{\frac{y^6(1-y)}{(1-ky)\left(y^6+(1-ky)\left(1-y^6\right)\right)}, \frac{y^6\left(1-y^2\right)}{(1-ky)\left(y^6+(1-ky)\left(1-y^6\right)\right)}, \frac{y^6}{(1-ky)\left(1+y^5+y^6-k\left(y+y^6\right)\right)}, \frac{y^6}{(1-ky)(1-ky+y^6)}\right\}$.

The $p(3)$ and $p(4)$ cases split up as follows;

For period $p(3)$, we have $\{\bar{f}_{abaaba}, \bar{f}_{aabaab} = \bar{f}_{abbabb} = \bar{f}_{abcabc}\}$, where from Eqs (28) and (27) $\{\bar{f}_{abaaba} = \frac{1+y^3+y^5}{(1+y^3+y^5)(1-ky)+y^6}, \bar{f}_{aabaab} = \frac{1+y^3}{(1+y^3)(1-ky)+y^6}\}$.

And for period $p(4)$, we have $\{\bar{f}_{aaabaa} = \bar{f}_{aabcaa}, \bar{f}_{abaaab} = \bar{f}_{abccab} = \bar{f}_{abcdab}\}$, where from Eqs (29) and (30), $\{\bar{f}_{aaabaa} = \frac{1+y^4+y^5}{1+y^4+y^5+y^6-k\left(y+y^5+y^6\right)}, \bar{f}_{abaaab} = \frac{1+y^4}{1+y^4+y^6-k\left(y+y^5\right)}\}$. Thus we have proved

Theorem 5. *The generating function for the number of distinct* 6*-tuples in words of length n tracked by y, over an alphabet* $[k]$ *is given by*

$$\begin{aligned} F_k^{(6)}(y) =& k f_{6,1}(y) + (k)_2 f_{6,2}(y) + \underbrace{(k)_2 f_{abaaba} + 2(k)_2 f_{aabaab} + (k)_3 f_{abcabc}}_{period\ 3} \\ &+ \underbrace{(k)_4 f_{abcdab} + 5(k)_3 f_{abccab} + (k)_3 f_{aabcaa}}_{period\ 4} + \underbrace{3(k)_2 f_{abaaab} + 3(k)_2 f_{aaabaa}}_{period\ 4} \\ &+ \left((k)_5 + 10(k)_4 + 24(k)_3 + 11(k)_2\right) f_{6,5}(y) \\ &+ \left((k)_6 + 14(k)_5 + 54(k)_4 + 59(k)_3 + 10(k)_2\right) f_{6,6}(y), \end{aligned} \tag{34}$$

and the particular 6*-tuple generating functions are given prior to the theorem.*

This simplifies to

$$\begin{aligned} F_k^{(6)}(y) =& k f_{6,1}(y) + \left(k^2 - k\right) f_{6,2}(y) + k(k-1) f_{abaaba} + (k^3 - k^2) f_{abcabc} \\ &+ (k^4 - 4k^2 + 3k) f_{abcdab} + 3k(k-1) f_{aaabaa} \\ &+ \left(k^5 - k^3 - k^2 + k\right) f_{6,5}(y) + \left(k^6 - k^5 - k^4 + k^2\right) f_{6,6}(y). \end{aligned}$$

7. The Number of Restricted Growth Functions of Period I with J Distinct Letters

In this section we use the standard notation $S(n, j)$ to mean the Stirling set number which counts the number of partitions of an n element set into j blocks. As explained in Section 4, this is in bijection with restricted growth functions of length n using j different letters. As a reminder to the reader, we

show a further example of how this works: Consider the restricted growth function 1213. This is a restricted growth function of length 4 using 3 different letters. It encodes the set partition of [4] with 3 blocks in which the first set (block) contains the first and third elements (i.e., the positions of the 1s), the second set contains 1 element (i.e., the position of the 2s) and the third set contains only the element 4 (i.e., the position of the 3s).

Now we consider r-tuples of period i, $1 \leq i \leq r$, in which j distinct letters are used, $1 \leq j \leq i$, *only* in the case where there is a unique generating function for the number of such r tuples irrespective of the particular subword with this period and number of letters. For example, we will explain by means of an inclusion-exclusion argument how the coefficients $6, 18, 9, 1$ mentioned in subsection 5.1 were obtained.

The first 6 corresponds to 5-tuples of period $i = 5$ with $j = 2$ distinct letters and satisfies the following equation which is explained after it:

$$\begin{aligned} &S(5,2) - S(4,2) - S(3,2) - S(2,2) + (S(2,2) + S(2,2) + 0) - 0 \\ &= S(5,2) - S(4,2) - S(3,2) + S(2,2) = 6. \end{aligned} \tag{35}$$

In the above equation, we begin with the set of all RGFs of length 5 with two distinct letters. Such RGFs must have period 2 up to 5. Then we remove the number of restricted growth functions with maximum period $2, 3$ and 4 from the number with maximum period 5. From these, by the inclusion-exclusion argument, we need to add back the number of all pairwise intersections $(S(2,2)+S(2,2)+0)$ and remove the number contained in the intersection of all three (0). For example, the pairwise intersection of RGFs of period 4 and 3 is the set $\{11211\}$ of cardinality 1. The first $S(2,2)$ follows because the intersection of periods 4 and 3 force the first two and last two elements of the tuple to be 11. This means that the third and fourth elements must be 12 whose cardinality is counted by $S(2,2) = 1$. The last $S(2,2)$ in the added back intersections of the above equation follows because any period 2 case is also a period 4 case, and 121212 is the only 6-tuple of period 2.

The next number 18 corresponds to 5-tuples of period $i = 5$ with $j = 3$ distinct letters and is obtained as follows;

$$S(5,3) - S(4,3) - S(3,3) + (0) - (0) = 18, \tag{36}$$

where the explanation is similar to the case of the previous equation.

As a different type of example, we focus on 6-tuples of period $i = 5$ with $j = 2$ distinct letters. This differs because the period is not equal to the tuple length and requires an alternative explanation: We start with RGFs of 6 with 2 distinct letters with the additional property of having period 5 or less. This set A corresponds to RGFs of 5 with 2 distinct letters with an additional letter 1 appended to each such RGF of 5. The cardinality of A is $S(5,2)$. From set A we remove the set A_1 of all RGFs that also have period 4 or less and the set A_2 of all RGFs that also have period 3 or less. The RGFs in the set A_1 must start with 11 and end with 11. The 1 in position 2 together with the entries in positions 3 and 4 make up an RGF of length 3 with 2 distinct letters, counted by $S(3,2)$. The RGFs in the set A_2 must have 1 in positions $1, 3$ 4 and 6, leaving a 2 in position 2 and position 5. Thus the first two entries make up an RGF of length 2 with two distinct letters counted by $S(2,2)$. Finally add back the size of $A_1 \cap A_2$ which is empty. this yields the equation

$$S(5,2) - S(3,2) - S(2,2) + 0 = 11. \tag{37}$$

Another example is 6-tuples of period $i = 6$ with $j = 2$ distinct letters. We obtain

$$\begin{aligned} &S(6,2) - S(5,2) - S(4,2) - S(3,2) - S(2,2) + (S(3,2) + S(2,2) + 0 + 0 + S(2,2) + 0) - 0 \\ &= S(6,2) - S(5,2) - S(4,2) + S(2,2) = 10. \end{aligned} \tag{38}$$

We begin with all RGFs of 6-tuples with 2 distinct letters, counted by $S(6,2)$ and remove from these the tuples of maximum period, $5, 4, 3$ and 2 of respective cardinalities $S(5,2)$, $S(4,2)$, $S(3,2)$

and $S(2,2)$. We then respectively add back the pairwise intersections of maximum periods 5 and 4, maximum periods 5 and 3, maximum periods 5 and 2, maximum periods 4 and 3, maximum periods 4 and 2 and finally maximum periods 3 and 2. There were no triple or higher intersections. The respective cardinalities of these intersections are $3 = S(3,2); 1 = S(2,2); 0; 0; 1 = S(2,2); 0$. The first 2 Stirling numbers have been explained in the previous example. The last $S(2,2)$ follows as before because any period 2 case is also a period 4 case, and 121212 is the only 6-tuple of period 2.

8. The Triangles of Coefficients

In the triangles below, each position represents either a single or multiple entry.

For the case where we have a single entry in row i and column j of each r-tuple table below, this is the coefficient of $k(k-1)...(k-j+1)$ in the factor multiplying $f_{r,i}(y)$, where $f_{r,i}(y)$ is the unique generating function for such r-tuples of period i. The number of letters used is counted by j. The multiple entries are similar but imply that there is not a unique generating function for such r- tuples. We first display the case of r-tuples, for $r = 5$, taking into account that we have an additional split that occurs in row 3 because there is not just a single generating function arising for period 3.

The table of coefficients for 5-tuples is

$$\begin{array}{ccccc} 1 & & & & \\ 0 & 1 & & & \\ 0 & \{1,2\} & 1 & & \\ 0 & 5 & 6 & 1 & \\ 0 & 6 & 18 & 9 & 1 \end{array}$$

where $\{1,2\}$ in the above table are the coefficients of $\{f_{aabaa}, f_{abbab}\}$ respectively. In this same row the unique entry 1 following $\{1,2\}$ is understood to be multiplying $f_{5,3}(y) := f_{abcab}(y)$.

The table of coefficients for 6-tuples is

$$\begin{array}{ccccccc} 1 & & & & & & \\ 0 & 1 & & & & & \\ 0 & \{1,2\} & 1 & & & & \\ 0 & \{3,3\} & \{1,5\} & 1 & & & \\ 0 & 11 & 24 & 10 & 1 & & \\ 0 & 10 & 59 & 54 & 14 & 1 & \end{array}$$

where $\{1,2\}$ in the above table are the coefficients of $\{f_{abaaba}, f_{aabaab}\}$ and $\{1,5\}$ are the coefficients of $\{f_{aaabaa}, f_{abccab}\}$ respectively. As in the previous table the 1 in position 4 of row 4 is multiplying $f_{6,4}(y) := f_{abcdab}(y)$.

9. Further Research

In this paper we restricted the size of the r-tuples to at most 6. This is because as the size of the r-tuple becomes larger, the complexity of the problem increases. The approach we used in the paper could be extended to slightly larger r but not arbitrary such values. The authors plan to consider a generalisation of this question to r-tuples of any size, and invite interested researchers to do likewise.

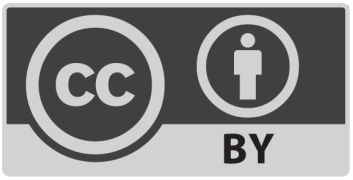
Conflict of Interest

The authors declare that they have no conflicts of interest.

References

1. Archibald, M., Blecher, A., Brennan, C., Knopfmacher, A., Wagner, S. and Ward, M., 2021. The number of distinct adjacent pairs in geometrically distributed words. *Discrete Mathematics & Theoretical Computer Science, 22*(4), 10, (Analysis of Algorithms).
2. Archibald, M., Blecher, A. and Knopfmacher, A., 2019. Distinct r-tuples in integer partitions. *The Ramanujan Journal, 50*, pp.237-252.

3. M. Hirschhorn, (2014). The number of different parts in the partitions of n. *The Fibonnacci Quarterly 52*(1), 10-15.
4. Archibald, M., Knopfmacher, A. and Prodinger, H., 2006. The number of distinct values in a geometrically distributed sample. *European Journal of Combinatorics, 27*(7), pp.1059-1081..
5. Bassino, F., Clément, J. and Nicodème, P., 2012. Counting occurrences for a finite set of words: combinatorial methods. *ACM Transactions on Algorithms (TALG), 8*(3), pp.1-28.
6. Mansour, T., 2013. Combinatorics of set partitions. Boca Raton: CRC Press.



© 2024 the Author(s), licensee Combinatorial Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)