TUCS

Tero Harju | Dirk Nowotka

Counting Bordered and Primitive Words with a Fixed Weight

Turku Centre *for* Computer Science

TUCS Technical Report
No 630, November 2004

Counting Bordered and Primitive Words with a Fixed Weight

Tero Harju
Turku Centre for Computer Science and Department of Mathematics,
University of Turku, FIN-20014 Turku, Finland
Dirk Nowotka
Institute of Formal Methods in Computer Science,
University of Stuttgart, D-70569 Stuttgart, Germany

TUCS Technical Report
No 630, November 2004

Abstract

A word w is primitive if it is not a proper power of another word, and w is unbordered if it has no prefix that is also a suffix of w. We study the number of primitive and unbordered words w with a fixed weight, that is, words for which the Parikh vector of w is a fixed vector. Moreover, we estimate the number of words that have a unique border.

Keywords: combinatorics on words, borders, primitive words

TUCS Laboratory
Discrete Mathematics for Information Technology

1 Introduction

Let w denote a finite word over some alphabet A. We say that w is bordered if there is a nonempty proper prefix x of w that is also a suffix of w. If there is no such x then w is called unbordered. We say that w is primitive if $w = x^k$, for some $k \in \mathbb{N}$, implies that $k = 1$ and $x = w$. We often assume that the alphabet is ordered, $A = \{a_1, a_2, \ldots, a_q\}$. In this case, for a word $w \in A^*$, let $\pi(w)$ denote by $(|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_q})$ the *Parikh vector* of w, where $|w|_a$ denotes the number of occurrences of the letter a in w. We also say that w has weight $\pi(w)$.

The number of primitive words and unbordered words of a fixed length and an alphabet of a fixed size is well-known, see for example [4, 3, 1, 2, 5, 6] and the sequences A027375, A003000, A019308, and A019309 in Sloane's database of integer sequences [7]. We will recall these results with short arguments and extend them to the case where the words we consider have a fixed weight. Moreover, we estimate the number of words that have exactly one border.

Section 2 contains results on counting the number of primitive words. Section 3 investigates the number of bordered words. Finally, we deal with the number of words with exactly one border in section 4. In the rest of this section we will fix our notation. For more general definitions see [2].

Let A be a finite, non-empty set called *alphabet*. The elements of A are called *letters*. Let a finite sequence of letters be called (finite) *word*. Let A^* denote the monoid of all finite words over A where ε denotes the *empty word*. Let $|w|$ denote the length of w, and let $|w|_a$ denote the number of occurrences of a in w where $a \in A$. If $w = uv$ then u is called *prefix* of w, denoted by $u \leq_{\mathrm{p}} w$, and v is called *suffix* of w, denoted by $v \leq_{\mathrm{s}} w$. A word w is called *bordered* if there exist non-empty words x, y, and z such that $w = xy = zx$, and x is called a *border* of w. Let X be a set, then $|X|$ denotes the cardinality of X.

The Möbius function $\mu\colon \mathbb{N} \to \mathbb{Z}$ is defined as follows

$$\mu(n) = \begin{cases} (-1)^t & \text{if } n = p_1 p_2 \ldots p_t \text{ for distinct primes } p_i, \\ 1 & \text{if } n = 1, \\ 0 & \text{if } n \text{ is divisible by a square.} \end{cases}$$

The Möbius inversion formula for two functions f and g is given by:

$$g(n) = \sum_{d|n} f(d)$$

if and only if

$$f(n) = \sum_{d|n} \mu(d)\, g(n/d)\ .$$

2 Primitive Words

Let $P_q(n)$ denote the number of primitive words of length n over an alphabet of size q. It is well-known, see for example [4, 2] and the sequence A027375 in [7], that

$$P_q(n) = \sum_{d|n} \mu(d)\ q^{n/d}\ . \tag{1}$$

Indeed, let A with $|A| = q$ be a finite alphabet of letters. Every word w has a unique primitive root v for which $w = v^d$ for some $d|n$, where $n = |w|$. Since there are exactly q^n words of length n,

$$q^n = \sum_{d|n} P_q(d)\ .$$

We are in the divisor poset, where the Möbius inversion gives (1).

In this paper we investigate the number of primitive words with a fixed weight, that is a fixed number of occurrences of letters. Consider an ordered alphabet $A = \{a_1, a_2, \ldots, a_q\}$ of $q \geq 1$ letters. For a word $w \in A^*$, let $\pi(w)$ denote $(|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_q})$ which is called the *Parikh vector* of w. For a given vector $\mathbf{k} = (k_1, k_2, \ldots, k_q)$, let

$$\mathcal{P}(\mathbf{k}) = \{w \mid w \text{ primitive and } \pi(w) = \mathbf{k}\}$$

and let $P(\mathbf{k}) = |\mathcal{P}(\mathbf{k})|$. Clearly, if $w \in \mathcal{P}(\mathbf{k})$, then $|w| = \sum_{i=1}^{q} k_i$. Also, denote by $\gcd(\mathbf{k})$ the greatest common divisor of the components k_i. If $d|\gcd(\mathbf{k})$, then denote

$$\mathbf{k}/d = (k_1/d, k_2/d, \ldots, k_q/d)\ .$$

The multinomial coefficients under consideration are

$$\binom{n}{\mathbf{k}} = \binom{n}{k_1, k_2, \ldots, k_q} = \frac{n!}{k_1!\, k_2!\, \ldots\, k_q!}\ ,$$

where $n = \sum_{i=1}^{q} k_i$.

Theorem 1. *Let $\mathbf{k} = (k_1, k_2, \ldots, k_q)$ be a vector with $n = \sum_{i=1}^{q} k_i$. Then*

$$P(\mathbf{k}) = \sum_{d|\gcd(\mathbf{k})} \mu(d) \binom{n/d}{\mathbf{k}/d}\ .$$

Proof. We use the principle of inclusion and exclusion to prove our claim. Let the distinct prime divisors of $\gcd(\mathbf{k})$ be $p_1, p_2, \ldots, p_t$.

For an integer $d|\gcd(\mathbf{k})$, define

$$Q_d = \{w \mid w = u^d \text{ where } \pi(u) = \mathbf{k}/d\}\ .$$

If $w \in Q_d$, then $\pi(w) = \mathbf{k}$. Clearly, $|Q_d|$ equals the number of all words u, primitive and imprimitive alike, of length n/d such that u has the Parikh vector $\mathbf{k}/d$. Therefore,

$$|Q_d| = \binom{n/d}{\mathbf{k}/d} \ . \tag{2}$$

Notice also that if $d|e$, then $Q_e \subseteq Q_d$, and hence

$$I(\mathbf{k}) = \bigcup_{i=1}^{t} Q_{p_i} \tag{3}$$

is the set of all imprimitive words of length n with Parikh vector $\mathbf{k}$. By the principle of inclusion and exclusion, we have then that

$$\Big|\bigcup_{i=1}^{t} Q_{p_i}\Big| = \sum_{\emptyset \neq Y \subseteq [1,t]} (-1)^{|Y|-1} \ \Big|\bigcap_{i \in Y} Q_{p_i}\Big|, \tag{4}$$

where $\bigcap_{i\in Y} Q_{p_i} = Q_{p(Y)}$ for $p(Y) = \prod_{i \in Y} p_i$. Hence, by (2),

$$\begin{aligned} |I(\mathbf{k})| &= \sum_{\emptyset \neq Y \subseteq [1,t]} (-1)^{|Y|-1} \ \Big|Q_{p(Y)}\Big| \\ &= - \sum_{\emptyset \neq Y \subseteq [1,t]} (-1)^{|Y|} \ \binom{n/p(Y)}{\mathbf{k}/p(Y)} \\ &= - \sum_{\substack{d|\,\gcd(\mathbf{k}) \\ d>1}} \mu(d) \binom{n/d}{\mathbf{k}/d}, \end{aligned}$$

by the definition of the Möbius function μ. This proves the claim, because $P(\mathbf{k}) = \binom{n}{\mathbf{k}} - |I(\mathbf{k})|$. □

3 Unbordered Words

Let $U_q(n)$ denote the number of all unbordered words of length n over an alphabet of size q. The following formula for $U_q(n)$ is well-known, see for example [3, 1, 5, 6] and also the sequences A003000, A019308, A019309 in [7]. Surely, we have $U_q(1) = q$ and if $n > 1$ then

$$U_q(2n+1) = q\, U_q(2n) \tag{5}$$

$$U_q(2n) = q\, U_q(2n-1) - U_q(n) \ . \tag{6}$$

Indeed, case (5) is clear since a word of odd length is unbordered if and only if it is unbordered after its middle letter (at position $n+1$) is deleted. For case (6) consider that a word w of even length is unbordered if and only if it is unbordered after one of its middle letters (say, at position $n+1$) is deleted except if $w = auau$ and au is unbordered, where a is an arbitrary letter.

Note, that there is an alternative way to obtain $U_q(n)$ by considering the following immediate result.

Lemma 2. *Let w be a bordered word, and let u be its shortest border. Then*

1. $2|u| \leq |w|$,
2. *u is unbordered, and*
3. *u is the only unbordered border of w.*

Let $B_q(n)$ denote the number of all bordered words of length n over an alphabet of size q. Lemma 2 shows that it is enough for every unbordered border u, with $|u| \leq \lfloor n/2 \rfloor$, to count the number of words of length $n - 2|u|$ which is $q^{n-2|u|}$. So, we have

$$B_q(n) = \sum_{1 \leq i \leq \lfloor n/2 \rfloor} U_q(i)\, q^{n-2i} \ .$$

This gives the following formula for $U_q(n)$ where

$$U_q(n) = q^n - B_q(n) \tag{7}$$

for every $q > 1$ and where $U_q(1) = q$.

In this paper we investigate the number of unbordered words with a fixed weight. Let us fix a binary alphabet $A = \{a, b\}$ for now. Let $U(n, k)$ denote the number of all binary unbordered words of length n that have a fixed weight k in the sense that, for every such word w, we have $|w|_b = k$ and $|w|_a = n - k$.

It is easy to check that $U(1, 0) = U(1, 1) = 1$ and $U(n, k) = 0$, if $n \leq k$ and $k > 1$, and $U(n, 0) = 0$, if $n > 1$.

Theorem 3. *If $0 < k < n$ then*

$$U(n,k) = U(n-1,k) + U(n-1,k-1) - E(n,k) \tag{8}$$

where

$$E(n,k) = \begin{cases} U(n/2, k/2) & \text{if } n \text{ and } k \text{ are even,} \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Suppose first that w has odd length $2n+1$. Each word $w = ucv$, with $c \in A$ and $|u| = |v| = n$, contributing to $U(2n+1, k)$ is obtained by adding a middle letter c to an unbordered word uv of even length. If $c = a$ then uv contributes to $U(2n, k)$, and if $c = b$ then uv contributes to $U(2n, k-1)$.

Assume then that w has even length $2n$. If $w = cudv$, with $c, d \in A$ and $|u| = |v| = n - 1$, then it contributes to $U(2n, k')$ if and only if cuv is unbordered (so it contributed to either $U(2n-1, k')$ or $U(2n-1, k'-1)$) and $cu \neq dv$ (that is, borderedness is not obtained by adding a letter to cuv such that w is a square). Consider the case where cuv is unbordered but $cudv$ is not, that is, $cu = dv$. Then $w = cucu$ and cuu is unbordered. Note, that cuu

is unbordered if and only if cu is unbordered. Let $|cu|_b = k$. We have that cuu contributes to $U(2n-1, 2k)$ (if $c = a$) or $U(2n-1, 2k-1)$ (if $c = b$) if and only if cu contributes to $U(n, k)$ which is therefore subtracted in case $|w|_b = 2k$. □

Equation (8) can be generalized to alphabets of arbitrary size q. For this, consider an ordered alphabet $\{a_1, a_2, \ldots, a_q\}$ of size q, and let $U(\mathbf{k})$ denote the number of all unbordered words w of length $n = \sum_{i=1}^{q} k_i$ that have a fixed weight $\pi(w) = \mathbf{k} = (k_1, k_2, \ldots, k_q)$. Moreover, let $\mathbf{k}[k_i - 1]$ denote $(k_1, \ldots, k_{i-1}, k_i - 1, k_{i+1}, \ldots, k_q)$.

If $k_j = 1$ and $k_i = 0$ for all components $i \neq j$ of $\mathbf{k}$, then only the letter a_j contributes to $U(\mathbf{k})$. Hence $U(\mathbf{k}) = 1$, if $\sum_{i=1}^{q} k_i = 1$ and $k_i \geq 0$ for all $1 \leq i \leq q$.

Theorem 4. *If $\sum_{i=1}^{q} k_i > 0$ then*

$$U(\mathbf{k}) = \left[\sum_{\substack{1 \leq i \leq q \\ k_i > 0}} U(\mathbf{k}[k_i - 1]) \right] - E(\mathbf{k})$$

where

$$E(\mathbf{k}) = \begin{cases} U(\mathbf{k}/2) & \text{if } k_i \text{ is even for all } 1 \leq i \leq q, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Indeed, the arguments of adding a letter at the point $\lceil |w|/2 \rceil$ of a word w are similar to those of Theorem 3. For the explanation of $E(\mathbf{k})$ we note that a bordered word (created by adding a middle letter) is a square $a_i u a_i u$, for some $1 \leq i \leq q$. Note that the length of w and the number of occurrences of every letter is even in that case. Now, w is only counted if $a_i u$ is unbordered, that is, if $a_i u$ contributes to $U(\mathbf{k}/2)$ which must be therefore subtracted. □

4 Words with a Unique Border

In this section we count the number of words that have one and only one border. Let us start with an obvious result which belongs to folklore.

Lemma 5. *Let w be a bordered word, and let u be its shortest border. If w has a border v with $|v| > |u|$ border, then $|v| \geq 2|u|$.*

Proof. Indeed, if, for the shortest border u, we have $|v| < 2|u|$ then u overlaps itself (since $u \leq_{\mathrm{p}} v$ and $u \leq_{\mathrm{s}} v$), and hence, u is bordered contradicting Lemma 2(2). □

In order to estimate the number of words with exactly one border, we make following two observations.

Lemma 6. *Let u be a fixed unbordered word of length s. Then the number of words of length r of the form $xuyux$ is the number of bordered words of length $r-2s$, that is, $B_q(r-2s)$.*

Indeed, every word of the form xyx produces exactly one word of the form $xuyux$, and the condition $xuyux = x'uy'ux'$ would imply that u is bordered; a contradiction.

Lemma 7. *Let u be a fixed unbordered word of length s. Then the number of words of length r of the form zuz is the number of words of length $(r-s)/2$.*

Indeed, each word z produces exactly one word of the form zuz, and the condition $zuz = z'uz'$ implies that $z = z'$.

Let $k \leq n$ and $B_q(n,k)$ denote the number of all words of length n over an alphabet of size q that have exactly one border of length k. It is clear that $B_q(1,k) = B_q(n,0) = 0$, for all $1 \leq n$ and $0 \leq k$, and $B_q(n,k) = 0$, if $n < 2k$, see Lemma 2(1).

Theorem 8. *If $1 \leq 2k \leq n$ then*

$$B_q(n,k) = U_q(k)\left(q^{n-2k} - W_q(n-2k,k) - E_q(n-2k,k)\right)$$

where

$$W_q(r,s) = \begin{cases} B_q(r-2s) & \textit{if } 2s < r, \\ 1 & \textit{if } 2s = r, \\ 0 & \textit{otherwise.} \end{cases}$$

and

$$E_q(r,s) = \begin{cases} q^{(r-s)/2} & \textit{if } s < r < 3s \textit{ and } r-s \textit{ even,} \\ 1 & \textit{if } s = r, \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. Indeed, following the argument of Lemma 2(2) we count all unbordered words of length k (that is $U_q(k)$) which are possible borders of a word of length n. For every such border we have to count the number of different combinations of letters for the rest of the $n-2k$ letters, that is q^{n-2k}. However, we have to exclude those cases where new borders are created. Given an unbordered border u of length k, we have the following cases for words with more than one border: $uxuyuxu$ and $uzuzu$, where $x, y, z \in A^*$. These two cases are taken care of by $W_q(r,s)$ and $E_q(r,s)$ where both terms equal 1 if u^4 and u^3 are counted; see also Lemma 6 and 7. Note that the latter case is included in the former one if and only if $|u| \leq |z|$ (where the "only if" part comes from the fact that u is unbordered, and hence, it does not overlap itself), therefore $r < 3s$ is required in $E_q(r,s)$. □

Clearly, the number $B_q(n)$ of words of length n over an alphabet of size q with exactly one border is the following.

$$B_q(n) = \sum_{1 \le i \le \lfloor n/2 \rfloor} B_q(n, i) \ .$$

References

[1] Heiko Harborth. Endliche $0-1$-Folgen mit gleichen Teilblöcken. *J. Reine Angew. Math.*, 271:139–154, 1974.

[2] M. Lothaire. *Combinatorics on words*, volume 17 of *Encyclopedia of Mathematics and its Applications*. Addison-Wesley Publishing Co., Reading, Mass., 1983.

[3] P. Tolstrup Nielsen. A note on bifix-free sequences. *IEEE Trans. Information Theory*, IT-19:704–706, 1973.

[4] H. Petersen. On the language of primitive words. *Theoret. Comput. Sci.*, 161(1-2):141–156, 1996.

[5] M. Régnier. Enumeration of bordered words. The language of the laughing cow. *RAIRO Inform. Théor. Appl.*, 26(4):303–317, 1992.

[6] I. Simon. String matching algorithms and automata. In *Results and trends in theoretical computer science (Graz, 1994)*, volume 812 of *Lecture Notes in Comput. Sci.*, pages 386–395. Springer, Berlin, 1994.

[7] N. J. A. Sloane. On-line encyclopedia of integer sequences. `http://www.research.att.com/~njas/sequences/`.

Turku
Centre *for*
Computer
Science

Lemminkäisenkatu 14 A, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematical Sciences

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration
- Institute of Information Systems Sciences

ISBN 952-12-1444-9
ISSN 1239-1891