Bordered Conjugates of Words over Large Alphabets

Tero Harju
University of Turku
harju@utu.fi

Dirk Nowotka
Universität Stuttgart
nowotka@fmi.uni-stuttgart.de

Submitted: Oct 23, 2008; Accepted: Nov 14, 2008; Published: Nov 24, 2008
Mathematics Subject Classification: 68R15

Abstract

The border correlation function attaches to every word w a binary word $\beta(w)$ of the same length where the ith letter tells whether the ith conjugate $w' = vu$ of $w = uv$ is bordered or not. Let $[u]$ denote the set of conjugates of the word w. We show that for a 3-letter alphabet A, the set of β-images equals $\beta(A^n) = B^* \setminus \big(\big[ab^{n-1}\big] \cup D \big)$ where $D = \{a^n\}$ if $n \in \{5, 7, 9, 10, 14, 17\}$, and otherwise $D = \emptyset$. Hence the number of β-images is $B_3^n = 2^n - n - m$, where $m = 1$ if $n \in \{5, 7, 9, 10, 14, 17\}$ and $m = 0$ otherwise.

Keywords: combinatorics on words, border correlation, binary words, square-free, cyclically square-free, Currie set,

1 Introduction

The border correlation function of a word was introduced by the present authors in [4], where the binary case was considered in detail. In this paper we consider the case for alphabets of size $s \geq 3$. The border correlation function is related to the *auto-correlation* function of Guibas and Odlyzko [3], as well as to the *border-array* function of Moore, Smyth and Miller [7]. Border correlation of partial words have been recently considered by Blanchet-Sadri et al. [1].

A word $w \in A^*$ is said to be *bordered* (or *self-correlated* [8]), if there exists a nonempty word v, with $v \neq w$, such that $w = u_1 v = v u_2$ for some words u_1, u_2. In this case v is a *border* of w. A word that has a border is called *bordered*; otherwise it is *unbordered*.

Let $\sigma\colon A^* \to A^*$ be the (cyclic) *shift function*, where $\sigma(xw) = wx$ for all $w \in A^*$ and $x \in A$, and $\sigma(\varepsilon) = \varepsilon$ for the empty word ε. Let $B = \{a, b\}$ be a special binary alphabet. The *border correlation function* $\beta\colon A^* \to B^*$ is defined as follows. For the empty word, let $\beta(\varepsilon) = \varepsilon$. For a word $w \in A^*$ of length n, let $\beta(w) = c_0 c_1 \ldots c_{n-1} \in B^*$ be the binary word of the same length such that

$$c_i = \begin{cases} a & \text{if } \sigma^i(w) \text{ is unbordered,} \\ b & \text{if } \sigma^i(w) \text{ is bordered.} \end{cases}$$

Example 1. (1) Assume the word w is not primitive, i.e., $w = u^k \, (= uu \dots u)$, for some power $k \geq 2$. Then all words $\sigma^i(w)$ are bordered, and thus $\beta(w) = b^n$, where n is the length of w.

(2) Consider the alphabet $A = \{a, b, c\}$, and let $w = bacaba \in A^*$. Then

i	$\sigma^i(w)$	border	i	$\sigma^i(w)$	border
0	$bacaba$	ba	3	$ababac$	-
1	$acabab$	-	4	$babaca$	-
2	$cababa$	-	5	$abacab$	ab

and hence $\beta(w) = baaaab$. Note that a border need not be unique.

For an alphabet A, let A^* denote the monoid of all finite words over A including the empty word ε. Also, let A^n denote the set of words $w \in A^*$ of length n. In the binary case, where we can choose $A = B \, (= \{a, b\})$, it was shown in [4] that the image $\beta(w)$ of $w \in B^*$ does not have two consecutive a's except for some trivial cases. Hence, if $\sigma^i(w)$ is unbordered, then $\sigma^{i+1}(w)$ is necessarily bordered. Also, in the binary case, there are other 'exceptions' , e.g., for no binary word w, it is the case that $\beta(w) = abababbababb$. It is an open problem to characterize the set of the images $\beta(w)$ for $w \in B^*$.

The words xy and yx are called *conjugates* of each other. We denote by $[w]$ the set of all conjugates of the word w. Note that if u and v are conjugates then $v = \sigma^i(u)$ for some i, and hence, for all words w,

$$\beta([w]) = [\beta(w)]\,. \tag{1}$$

Let $\beta(A^n) = \{\beta(w) \mid w \in A^n\}$ be the set of the β-*images* of the words of length n, and denote by B_k^n the cardinality of $\beta(A^n)$ where A is a k-letter alphabet. In the present paper we prove the following result, where

$$\mathbf{C} = \{5, 7, 9, 10, 14, 17\}$$

is the *Currie set* of integers.

Theorem 1. *Let A be an alphabet of three letters, and let $n \geq 2$. Then*

$$\beta(A^n) = \begin{cases} B^* \setminus [ab^{n-1}] & \text{if } n \notin \mathbf{C}, \\ B^* \setminus ([ab^{n-1}] \cup \{a^n\}) & \text{if } n \in \mathbf{C}. \end{cases}$$

In particular, $B_3^n = 2^n - n - m$, where $m = 1$ if $n \in \mathbf{C}$ and $m = 0$ otherwise.

We end this section with some definitions and notation needed in the rest of the paper. We refer to Lothaire's book [6] for more basic and general definitions of combinatorics on words.

We denote the length of a word w by $|w|$. A word u is a *factor* of a word $w \in A^*$, if $w = w_1 u w_2$ for some words $w_1 \in A^*$ and $w_2 \in A^*$. A word $w \in A^*$ is said to be *square-free*, if it does not have a factor of the form vv where $v \in A^*$ is nonempty. Moreover, w is *cyclically square-free*, if all its conjugates are square-free.

2 The proof

This section let $A = \{a, b, c\}$ be a ternary alphabet. Let T denote the *Thue word* obtained by iterating the substitution $\varphi\colon \{a, b, c\}^* \to \{a, b, c\}^*$ determined by $\varphi(a) = abc$, $\varphi(b) = ac$ and $\varphi(c) = b$. Therefore T is the infinite word starting with

$$T = abcacbabcbacabcacbacabcba\ldots$$

As was shown by Thue [9, 10] (see also Lothaire [5]), the word T is square-free, i.e., it does not contain any nonempty factors of the form vv.

Recall that $[w]$ denotes the conjugacy class of the word w. By the next lemma, each primitive word has at least two unbordered conjugates.

Lemma 1. *For all $n \geq 2$, $[ab^{n-1}] \cap \beta(A^n) = \emptyset$.*

Proof. Assume a occurs in $\beta(w)$ for a word w with $|w| \geq 2$. Hence w is primitive. A conjugate v of w is a *Lyndon word* if it is minimal in $[w]$ with respect to some lexicographic order of A^*. It is well known (see, e.g., Lothaire [6]), that each primitive word w has a unique Lyndon conjugate with respect to a given order and that each Lyndon word is unbordered. Hence, there exists at least two Lyndon words in $[w]$ for a given order of A and its inverse order, respectively. These two words imply that a occurs at least twice in $\beta(w)$. □

The following result is due to Currie [2].

Theorem 2 (Currie). *There exists a cyclically square-free word $w \in A^n$, if and only if $n \notin \mathbf{C} = \{5, 7, 9, 10, 14, 17\}$.*

A square vv is called *simple* if $v \in a^*$ with $v \neq \varepsilon$. Let $w_{(i)}$ denote the i-th letter of w.

Lemma 2. *Let w be a square-free word. Then $w' = w_{(1)}^{k_1} w_{(2)}^{k_2} \cdots w_{(n)}^{k_n}$ contains only simple squares for all $1 \leq i \leq n$ and $k_i \geq 1$.*

Proof. Suppose on the contrary that w' contains a nonsimple square vv, say

$$\begin{aligned} v &= b_{i+1}^{p_{i+1}} b_{i+2}^{p_{i+2}} \cdots b_{i+j-1}^{p_{i+j-1}} b_{i+j}^{p_{i+j}} \\ &= b_{i+j+1}^{p_{i+j+1}} b_{i+j+2}^{p_{i+j+2}} \cdots b_{i+2j-1}^{p_{i+2j-1}} b_{i+2j}^{p_{i+2j}} \end{aligned}$$

with $0 \leq i \leq n-2j$ and $p_{i+1} \leq k_{i+1}$ and $p_{i+\ell} = k_{i+\ell} = k_{i+j+\ell-1}$, for all $2 \leq \ell < j$, and $p_{i+j} + p_{i+j+1} = k_{i+j}$ and $p_{i+j} \leq k_{i+2j-1}$ and $b_{i+1} = b_{i+j} = b_{i+2j} = w_{(i+j)} = w_{(i+2j-1)}$ and $b_{i+\ell} = b_{i+j+\ell} = w_{(i+\ell)} = w_{(i+j+\ell-1)}$, for all $1 \leq \ell < j$.

Observe that we obtain a square $(b_{i+1}b_{i+2}\cdots b_{i+j-1})^2$ from vv when all powers in vv are reduced to 1 and the last letter is deleted. But now, we have that $b_{i+1}b_{i+2}\cdots b_{i+j-1} = w_{(i+1)}w_{(i+2)}\cdots w_{(i+j-1)} = w_{(i+j)}w_{(i+j+1)}\cdots w_{(i+2j-2)}$ implies a square in w; a contradiction. $\square$

Lemma 3. *Let w be a cyclically square-free word of length $n \geq 2$. Then for each nonempty $u \in \{a,b\}^*$ that has exactly n occurrences of a, there exists a word w' such that $\beta(w') = u$.*

Proof. By (1), we can assume without loss of generality that u begins with the letter a. Let $u = ab^{k_1}ab^{k_2}\cdots ab^{k_n}$ where $k_i \geq 0$, for all $1 \leq i \leq n$. By Lemma 2, $w' = w_{(1)}^{k_1+1}w_{(2)}^{k_2+1}\cdots w_{(n)}^{k_n+1}$ and all its conjugates contain only simple squares. That is, if a conjugate $w_{(i)}^{k_i+1}w_{(i+1)}^{k_{i+1}+1}\cdots w_{(n)}^{k_n+1}w_{(1)}^{k_1+1}\cdots w_{(i-1)}^{k_{i-1}+1}$ of w' that starts and ends in different letters is bordered then $w_{(i)}w_{(i+1)}\cdots w_{(n)}w_{(1)}\cdots w_{(i-1)}$ is bordered contradicting the fact that w is cyclically square-free. This means that every conjugate of w' that starts and ends in a different letter is unbordered and all other conjugates are, of course, bordered by a border of length one. Hence, we have $\beta(w') = u$ which completes the proof. $\square$

Lemma 4. *Let $n \in \mathbf{C}$. Then $u = ab^{k_1}ab^{k_2}\cdots ab^{k_n} \in \beta(A^*)$ whenever $u \notin a^*$.*

Proof. Consider the following six words with lengths in $\mathbf{C}$ which have a unique border v of length two or three (the borders are underlined):

$$
\begin{aligned}
5&: \underline{ab}c\underline{ab}\\
7&: \underline{abc}b\underline{abc}\\
9&: \underline{ab}cacbc\underline{ab}\\
10&: \underline{ab}cacbac\underline{ab}\\
14&: \underline{abc}bacabacb\underline{abc}\\
17&: \underline{ab}cabacbcabcbac\underline{ab}
\end{aligned}
$$

It is straightforward to check that for every word w in the list, each $x \in [w]$ with $x \neq w$ is unbordered, i.e., there exists only one bordered word w in the conjugacy class $[w]$ and w has a unique border. This also implies that these words are square-free.

Let

$$u = ab^{k_1}ab^{k_2}\cdots ab^{k_n}$$

as in the statement of the lemma.

We proceed by case distinction on $|v|$ to show that for every n there exists a word w' such that $\beta(w') = u$ except if $k_1 = k_2 = \cdots = k_n$ for n equal to 5, 7, 9, 14, or 17, and $k_1 = k_3 = k_5 = k_7 = k_9$ and $k_2 = k_4 = k_6 = k_8 = k_{10}$ for $n = 10$. The exceptional cases are handled at the end of the proof.

Let $w \in A^*$ be any square-free word having a unique border v such that each word in $[w] \setminus \{w\}$ is unbordered. Write $w = w_{(1)}w_{(2)} \dots w_{(n)}$, where again $w_{(i)}$ denotes the ith letter of w.

Suppose first that $|v| = 3$ as in the case for 7 and 14. We can assume that $v = abc$ (possibly by renaming the letters); otherwise v would not be a unique border. Hence $w_{(1)}w_{(2)}w_{(3)} = abc = w_{(n-2)}w_{(n-1)}w_{(n)}$. Consider $w' = w_{(1)}^{k_1+1}w_{(2)}^{k_2+1} \cdots w_{(n)}^{k_n+1}$. Since exactly one conjugate of w is bordered, the number of the letter a in the β-image equals n, if w' is unbordered. Now, w' is unbordered if $k_2 \neq k_{n-1}$, and in this case $\beta(w') = u$. Note that, by (1), it is enough to show that $\beta(w') = u'$ for any conjugate u' of u. In particular, we are done if the powers k_i can be cycled so that, for some j, the word $w'' = w_{(1)}^{k'_1+1}w_{(2)}^{k'_2+1} \cdots w_{(n)}^{k'_n+1}$, where $k'_i = k_{i+j \bmod n}$, is unbordered. It follows that, for the border length 3, the only cases left in $n \in \mathbf{C}$ are when $k_1 = k_2 = \cdots = k_n$. (Note that the case $n = 9$, where n is divisible by 3, is treated below.)

Suppose then that $|v| = 2$ as in the case for 5, 9, 10, and 17. We can assume that $v = ab$ (possibly after renaming of the letters), i.e., $w_{(1)}w_{(2)} = ab = w_{(n-1)}w_{(n)}$. Consider $w' = w_{(1)}^{k_1+1}w_{(2)}^{k_2+1} \cdots w_{(n)}^{k_n+1}$. We recall that w is the unique bordered word in its conjugacy class. Now, w' is unbordered if $k_1 > k_{n-1}$ or $k_2 < k_n$. Analogously to the above case with $|v| = 3$ we can consider shifts of the indices modulo n. We conclude that w' is bordered for all possible shifts of $k_1, k_2, \dots, k_n$ only if $k_1 = k_2 = \cdots = k_n$ or n is even; a case that is avoided for $|v| = 2$ except for $n = 10$. If $n = 10$ then we are left with the case where $k_1 = k_3 = \cdots = k_9$ and $k_2 = k_4 = \cdots = k_{10}$, where possibly $k_1 = k_2$.

It remains to be shown that u is a β-image if $k_1 = k_2 = \cdots = k_n$ or $k_1 = k_3 = \cdots = k_9$ and $k_2 = k_4 = \cdots = k_n$, if $n = 10$, with $k_i \geq 1$ for all $1 \leq i \leq n$. Let $t = k_1 + 1$ and $s = k_2 + 1$. The following list gives a word for every $n \in \mathbf{C}$ such that the β-image is $(ab^{t-1})^n$ or $(ab^{t-1}ab^{s-1})^5$ in the case $n = 10$.

$$
\begin{aligned}
5&\colon a^tb^tc^ta^tbc^{t-1}\\
7&\colon a^tb^tc^tb^ta^tb^tcb^{t-1}\\
9&\colon a^tc^tb^ta^tb^tc^tb^ta^tcb^{t-1}\\
10&\colon c^tb^sa^tc^sa^tb^sc^ta^sc^tba^{s-1}\\
14&\colon b^tc^tb^ta^tb^tc^ta^tb^ta^tc^ta^tb^tc^tb^{t-1}a\\
17&\colon c^ta^tb^tc^ta^tc^tb^ta^tb^tc^tb^ta^tc^ta^tb^tc^tab^{t-1}
\end{aligned}
$$

This last claim can easily be verified by hand after noting that $s, t > 1$. This concludes the proof. $\square$

We now show that almost all binary words of length n are β-images.

Proof of the main Theorem 1. Let $u \in \{a, b\}^*$ be a nonempty binary word of length n. We proceed by a case distinction on the number k_a of occurrences of the letter a in u. Note that $\beta(a^n) = b^n$ for the case $k_a = 0$ and the case $k_a = 1$ does not exist; see Lemma 1.

Suppose $k_a \geq 2$. If $k_a \notin \mathbf{C}$ then there exists a cyclically square-free word w in A^* of length k_a by Theorem 2, and Lemma 3 shows how to construct a word w' such that $\beta(w') = u$.

In the remaining case, where $k_a \in \mathbf{C}$, we have $a^n \notin \beta(A^n)$ which explains the value of m; otherwise a cyclically square-free word of length $n \in \mathbf{C}$ would contradict Theorem 2. Lemma 4 shows that u is a β-image in the remaining cases.

Finally, by counting, we obtain the number of β-images: $B_3^n = 2^n - n - m$, where $m = 1$ if $n \in \mathbf{C}$ and $m = 0$ otherwise. $\square$

3 The case of four and more letters

The exceptions in the Currie set disappear when the alphabet has at least four letters.

Theorem 3. *$B_k^n = 2^n - n$ for all $k > 3$ and $n \geq 2$.*

Proof. It is sufficient to prove the claim for the alphabet of four letters, $A = \{a, b, c, d\}$, since $B_4^n = 2^n - n$ implies $B_k^n = 2^n - n$ for all $k > 3$. The n exceptions are the binary words of length n with only one letter a; see Lemma 1. We show that any binary word u of length n, except ab^{n-1} and its conjugates, is the β-image of a word over A. Note that $\beta(a^n) = b^n$. Let then $u \notin [ab^{n-1}]$, and suppose u has $k_a = m \geq 2$ occurrences of a. Let w be the prefix of the square-free Thue word T of length m where the last letter is replaced by d, that is, $w = vd$, where v is the prefix of T of length $m - 1$. Note that w is cyclically square-free because no square occurs in the prefix v, and no square can contain the letter d, since d occurs only once in u. Now, Lemma 3 implies the claim. $\square$

Acknowledgement

We are grateful to the anonymous referee of this journal for pointing out the second exception of the case $n = 10$ in the proof of Lemma 4.

References

[1] F. Blanchet-Sadri, E. Clader, and O. Simpson. Border correlations of partial words. *Theory Comput. Syst.* to appear.

[2] J. D. Currie. There are ternary circular square-free words of length n for $n \geq 18$. *Electron. J. Combin.*, 9(1):Note 10, 7 pp. (electronic), 2002.

[3] L. J. Guibas and A. Odlyzko. String overlaps, pattern matching, and nontransitive games. *J. Combin. Theory Ser. A*, 30(2):183–203, 1981.

[4] T. Harju and D. Nowotka. Border correlation of binary words. *J. Combin. Theory Ser. A*, 108(2):331–341, 2004.

[5] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics*. Addison-Wesley, Reading, MA, 1983.

[6] M. Lothaire. *Algebraic Combinatorics on Words*, volume 90 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, United Kingdom, 2002.

[7] D. Moore, W. F. Smyth, and D. Miller. Counting distinct strings. *Algorithmica*, 23(1):1–13, 1999.

[8] H. Morita, A. J. van Wijngaarden, and A. J. Han Vinck. On the construction of maximal prefix-synchronized codes. *IEEE Trans. Inform. Theory*, 42:2158–2166, 1996.

[9] A. Thue. Über unendliche Zeichenreihen. *Det Kongelige Norske Videnskabersselskabs Skrifter, I Mat.-nat. Kl. Christiania*, 7:1–22, 1906.

[10] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Det Kongelige Norske Videnskabersselskabs Skrifter, I Mat.-nat. Kl. Christiania*, 1:1–67, 1912.