Words with simple Burrows-Wheeler Transforms

Jamie Simpson

Department of Mathematics and Statistics
Curtin University of Technology, Perth, WA 6850, Australia
simpson@maths.curtin.edu.au

and

Simon J. Puglisi

School of Computer Science and Information Technology
RMIT University, Melbourne, Victoria 3001,Australia
sjp@cs.rmit.edu.au

Submitted: Mar 17, 2008; Accepted: Jun 9, 2008; Published: Jun 20, 2008
Mathematics Subject Classifications: 68R15, 68W05

Abstract

Mantaci et al have shown that if a word x on the alphabet $\{a, b\}$ has a Burrows-Wheeler Transform of the form $b^i a^j$ then x is a conjugate or a power of a conjugate of a standard word. We give an alternative proof of this result and describe words on the alphabet $\{a, b, c\}$ whose transforms have the form $c^i b^j a^k$. These words have some common properties with standard words. We also present some results about words on larger alphabets having similar properties.

1 Introduction

We use the usual notation for combinatorics on words. A word of n elements is $x = x[1..n]$, with $x[i]$ being the ith element and $x[i..j]$ the *factor* of elements from position i to position j. The letters in x come from some *alphabet* A. The set of all words with letters from A is A^*. The *length* of x, written $|x|$, is the number of letters in x and the number of occurrences of the letter a in x is $|x|_a$. A factor of length n is an *n-factor*. Two or more adjacent identical factors form a *power*. A word which is not a power is *primitive*. A word x or factor x is *periodic* with period p if $x[i] = x[i + p]$ for all i such that $x[i]$ and $x[i + p]$ are in the word. Two words x and y are *conjugate* if there exist words u and v such that $x = uv$ and $y = vu$. We write $C(x)$ for the set of conjugates of a word x. If x precedes y lexicographically we write $x \prec y$ and $x \preceq y$ means that either $x \prec y$ or $x = y$. Often we will use capital letters for sets of words. $X \prec Y$ means every word in the set X precedes every word in the set Y. If u and v are words then uXv is a set of words each having prefix u and suffix v. For any non-empty word x, $F(x)$ and $L(x)$ are, respectively, the first and last letters in x. If $x = \alpha_1\alpha_2\dots\alpha_n$ then the *reverse* of x, written x^R, is $\alpha_n\dots\alpha_2\alpha_1$. A word x is a *palindrome* if $x = x^R$. If m and n are integers we write $\gcd(m, n)$ for the greatest common divisor of m and n.

The Burrows-Wheeler Transform (henceforth BW Transform) [2] was introduced in 1994 as part of a data compression scheme, and has since been heavily studied (see, for example, [7] and [9] and references therein). To perform the transform on a word x first list its conjugates in lexicographic order. The transform is then formed by concatenating the final letters of the conjugates in this order. For example to transform "hello" we produce the list

$$elloh,\ hello,\ llohe,\ lohel,\ ohell$$

and obtain the transform *hoell*. We will write $BWT(x)$ for the BW Transform of x. The advantage of the transform is that for some words, such as English text, it produces transforms with many repeated letters, and these locally skew first-order statistics can be exploited by a compressor. For example the first two sentences of this paragraph transform to:

```
 no]mnhe.rW)fn4asaxsdstttmcsnmead mser [991  .   B- 2t  rr
rrpw dgiunsi  er rohhmchcehhldptrlo ssrseuotTtWc(phx   sfl
 nen rrrreoiiooeoaaaiTr cc icw fffffr nameeTtTgoopeooootru
 iaoteaaw nnnseirssraar  nidjBn o e
```

An extreme example of this is when all occurrences of each letter make up a factor in the transform. For example $BWT(bbabbba) = b^5a^2$ and $BWT(cacbca) = c^3ba^2$. It is interesting to ask what words have such BW Transforms: they represent the best case for BWT based compressors. In the case of words on a 2 letter alphabet this question was answered by Mantaci et al [6] who obtained the remarkable result that if $BWT(x) = b^ia^j$ then x is a conjugate or a power of a conjugate of a *standard word* (defined below). Standard words are, in a sense, the building blocks of the ubiquitous Sturmian words. It is surprising that they should turn up in connection with BW Transforms. It is not possible to have $BWT(x) = a^ib^j$ with i and j positive and a and b having their usual lexicographic order. In the next section we give a new proof of the Mantaci et al result, and in the third section obtain a similar result for words on a three letter alphabet. In the final section we present some results about words on larger alphabets having similar properties and compare our words on the three letter alphabet with standard words.

2 Size 2 alphabet

We consider words defined on the alphabet $A = \{a, b\}$. We will describe the set of all words on this alphabet which have BW Transforms of the form b^ia^j where i and j are non-negative integers. The main result of this section is Theorem 2.5 in which we describe such words with $\gcd(i,j)=1$. In Corollary 2.6 we use some results from [6] to obtain the general case. The morphisms ϕ and $\tilde{\phi}$ are defined by

$$\begin{array}{ll} \phi(a)=a & \tilde{\phi}(a)=ba \\ \phi(b)=ab & \tilde{\phi}(b)=b. \end{array}$$

We let S be the smallest set containing a and b which is closed under both ϕ and $\tilde{\phi}$. The set S is the set of *standard words* on $\{a,b\}$. See [5], Chapter 2, where standard words are defined in terms of ordered pairs (u,v) of standard words, each pair giving rise to the two ordered pairs (u,uv) and (vu,v). In our case the ordered pairs have the form $(X(a),X(b))$ where $X\in\{\phi,\tilde{\phi}\}^*$ and concatenation implies composition. The children of $(X(a),X(b))$ are $(X(a),X(a)X(b))$ and $(X(b)X(a),X(b))$ which equal $(X(\phi(a)),X(\phi(b)))$ and $(X(\tilde{\phi}(a)),\ X(\tilde{\phi}(b)))$ respectively. From this it is easy to see the equivalence of the definitions.

We will need the following lemmas. The first two are Propositions 2 and 3 from [6].

Lemma 2.1. *Two words x and y are conjugate if and only if $BWT(x)=BWT(y)$.*

Lemma 2.2. *If $x=u^d$ and $BWT(u)=\alpha_1\alpha_2\dots\alpha_n$ then $BWT(x)=\alpha_1^d\alpha_2^d\dots\alpha_n^d$.*

Lemma 2.3. *If x and y come from $\{a,b\}^*$, have the same length and $x\prec y$ then $\phi(x)\prec\phi(y)$ and $\tilde{\phi}(x)\prec\tilde{\phi}(y)$.*

Proof. Write $x=pas$ and $y=pbt$ for possibly empty strings p, s and t. Applying ϕ we have $\phi(x)=\phi(p)a\phi(s)$ and $\phi(y)=\phi(p)ab\phi(t)$. By the definition of ϕ, $\phi(s)$ must begin with an a so that $\phi(s)\prec b\phi(t)$ and so $\phi(x)\prec\phi(y)$. The proof of the second part is similar. □

Lemma 2.4. *If $x\in\{a,b\}^*$ is a conjugate of y then $\phi(x)$ is a conjugate of $\phi(y)$.*

Proof. By observation. □

We notice that in general $\{\phi(y) : y\in C(x)\}\neq C(\phi(x))$, as $\phi(x)$, being longer than x, has more conjugates.

Theorem 2.5. *$BWT(x)=b^ia^j$ for some i and j with $\gcd(i,j)=1$ if and only if x is a conjugate of a word in S.*

Proof. We first use induction on the length of x to show that every x in S has $BWT(x)$ of the required form, then show that the members of S are the only words with this property. It is clear that a and b belong to S and have BW Transforms of the appropriate form, so the statement holds for $|x|=1$.

Suppose that any $x\in S$ with $|x|<n$ has a BW Transform of the required form. Each member of $S\backslash\{a,b\}$ is the image under ϕ or $\tilde{\phi}$ of some other member. Consider a word y

in S with $|y| = n$. Then there exists x in S such that $\phi(x) = y$ or $\tilde{\phi}(x) = y$. Without loss of generality suppose $y = \phi(x)$. The conjugates of x make up sets

$$aX_1a,\ aX_2b,\ bX_3a,\ bX_4b.$$

Either aX_1a or bX_4b must be empty else $BWT(x)$ does not have the required form. Suppose that aX_1a is empty. The conjugates are lexicographically ordered thus:

$$aX_2b \prec bX_4b \prec bX_3a.$$

Applying ϕ to these and using Lemma 2.3 we get

$$a\phi(X_2)ab \prec ab\phi(X_4)ab \prec ab\phi(X_3)a.$$

The set of conjugates of $y = \phi(x)$ also includes $ba\phi(X_2)a$ and $bab\phi(X_4)a$. Since each member of $\phi(X_2)$ begins with a, the full set of conjugates, ordered lexicographically, is

$$a\phi(X_2)ab \prec ab\phi(X_4)ab \prec ab\phi(X_3)a \prec ba\phi(X_2)a \prec bab\phi(X_4)a.$$

By inspecting the final letters of each set we see that $BWT(y)$ has the form b^ia^j. A similar analysis applies if $y = \tilde{\phi}(x)$. By assumption $BWT(x) = b^{i'}a^{j'}$ for some i' and j' with $\gcd(i', j') = 1$. Then $i = i' + j'$ and $j = j'$ so that $\gcd(i, j) = 1$ as required. We have shown that any member of S has a BW Transform of the required form. We now show that the only words with such BW Transforms are conjugates of words in S.

By Lemma 2.1 words are conjugates if and only if they have the same BW Transform, so it is sufficient to show that all words of the form b^ia^j with $\gcd(i, j) = 1$ are transforms of some member of S. This is equivalent to showing that for all such i and j there is a member x of S with $|x|_a = i$ and $|x|_b = j$. This is proved easily by induction on $i + j$. It clearly holds when $i + j = 1$ since a and b are in S. Suppose it holds for all pairs (i, j) with $\gcd(i, j) = 1$ and $i+j < k$. Consider a pair (i', j') with $\gcd(i', j') = 1$ and $i'+j' = k$. Suppose $i' > j' \geq 1$. Then $\gcd(j', i' - j') = 1$ so S contains y, say, with $|y|_a = i' - j'$ and $|y|_b = j'$. But then a appears i' times in $\phi(y)$ and b appears j' times, as required. If $j' > i'$ the same reasoning applies with ϕ replaced by $\tilde{\phi}$. This completes the proof. □

Corollary 2.6. *A word x has BW Transform b^ia^j if and only if it is a conjugate of a word in S or a conjugate of a power of a word in S.*

Proof. We first show that for any i and j there is a word in S with BW Transform b^ia^j. Let $\gcd(i, j) = d$. If $d = 1$ then the statement is equivalent to the theorem. Otherwise write $i = pd$ and $j = qd$ where $\gcd(p, q) = 1$. By the theorem there exists x in S with $BWT(x) = b^pa^q$ and by Lemma 2.2 $BWT(x^d) = b^{pd}a^{qd}$ as required. The converse follows from Lemma 2.1. □

3 Size 3 alphabet

We now describe the set of words x on the alphabet $\{a, b, c\}$ with the property that

$$BWT(x) = c^i b^j a^k \tag{3.1}$$

for non-negative integers i, j and k. We call a word satisfying (3.1) a *Type I word.* Examples are given at the beginning of the last section. We will construct a set T of primitive words, each satisfying (3.1) and such that any primitive word satisfying (3.1) is a conjugate of a word in T. Then by Lemma 2.1 and Lemma 2.2 any Type I word is either a conjugate of a word in T or a power of such a conjugate.

The words in the set S of the last section satisfy (3.1) with $i = 0$. Let γ_1 be the morphism defined by

$$\gamma_1(a) = b,\ \gamma_1(b) = c.$$

It is easy to see that if $x \in S$ and $BWT(x) = b^j a^k$ then $BWT(\gamma_1(x)) = c^j b^k$. Similarly if γ_2 is defined by

$$\gamma_2(a) = a,\ \gamma_2(b) = c$$

then $BWT(\gamma_2(x)) = c^j a^k$, so both $\gamma_1(x)$ and $\gamma_2(x)$ are Type I. Let

$$T_0 = S \cup \{\gamma_1(x) : x \in S\} \cup \{\gamma_2(x) : x \in S\}. \tag{3.2}$$

The conjugates of the words in T_0 are the only primitive words which contain at most 2 distinct letters from $\{a, b, c\}$ and which satisfy (3.1).

We now extend the morphism ϕ defined in the last section to

$$\phi(a) = a,\ \phi(b) = ab,\ \phi(c) = ac.$$

Note that this agrees with the earlier definition when applied to a or b. We also need θ defined by

$$\theta(a) = c,\ \theta(b) = b,\ \theta(c) = a.$$

We will introduce a third mapping ψ below. We define T to be the minimal set of words which includes T_0 and is closed under the mappings θ, ϕ and ψ. To prove that T has the required properties we need several lemmas.

Let $x_1 \prec x_2 \prec \cdots \prec x_n$ be the conjugates of a word x having length n so that $BWT(x) = L[x_1]L[x_2]\ldots L[x_n]$. It is clear that the set of 2-factors occurring in x is precisely $\{L[x_i]F[x_i] : i = 1 \ldots n\}$. It is also clear that a necessary and sufficient condition for x to be Type I is that

$$x_i \prec x_j \Rightarrow L(x_i) \succeq L(x_j). \tag{3.3}$$

Lemma 3.1. *Let x be a Type I word with $|x|_a = \alpha$, $|x|_b = \beta$ and $|x|_c = \gamma$.*
(i) If $\beta + \gamma > \alpha \geq \gamma$ then the set of 2-factors in x is a subset of $\{ab, ac, ba, bb, ca\}$.
(ii) If $\alpha \geq \beta + \gamma$ then the set of 2-factors in x is a subset of $\{aa, ab, ac, ba, ca\}$.
(iii) If $\alpha + \beta > \gamma \geq \alpha$ then the set of 2-factors in x is a subset of $\{ac, bb, bc, ca, cb\}$.
(iv) If $\gamma \geq \alpha + \beta$ then the set of 2-factors in x is a subset of $\{ac, bc, ca, cb, cc\}$.

Proof. Let the conjugates of x be $x_1 \prec \cdots \prec x_n$. Since x is Type I $BWT(x) = c^\gamma b^\beta a^\alpha$ which is the concatenation $L(x_1)\ldots L(x_n)$. We also have $F(x_1)\ldots F(x_n) = a^\alpha b^\beta c^\gamma$. Consider the case $\beta+\gamma > \alpha \geq \gamma$. We see that

$$\begin{aligned}
F[x_i]L[x_i] = ac \text{ for } i &= 1\ldots\gamma \\
F[x_i]L[x_i] = ab \text{ for } i &= \gamma+1\ldots\alpha \\
F[x_i]L[x_i] = bb \text{ for } i &= \alpha+1\ldots\beta+\gamma \\
F[x_i]L[x_i] = ba \text{ for } i &= \beta+\gamma+1\ldots\alpha+\beta \\
F[x_i]L[x_i] = ca \text{ for } i &= \alpha+\beta+1\ldots\alpha+\beta+\gamma
\end{aligned}$$

Since the set of 2-factors in x is precisely the set of $L[x_i]F[x_i]$ values, part (i) of the Lemma follows. The proofs of the other parts are similar. □

Lemma 3.2. *Let x and y be words on the alphabet $\{a,b,c\}$.*
(a) If $x \prec y$ then $\phi(x) \prec \phi(y)$ and $\theta(x) \succ \theta(y)$.
(b) If x is a conjugate of y then $\phi(x)$ is a conjugate of $\phi(y)$ and $\theta(x)$ is a conjugate of $\theta(y)$.

Proof. (a) This is immediate since for any letters α and β from $\{a,b,c\}$ $\alpha \prec \beta$ implies $\phi(\alpha) \prec \phi(\beta)$ and $\theta(\alpha) \succ \theta(\beta)$.
(b) This is also immediate. □

Note that $\{\phi(y) : y \in C(x)\}$ includes all conjugates of $\phi(x)$ except those with prefix ba or ca and that $\{\theta(y) : y \in C(x)\}$ includes all conjugates of $\theta(x)$.

Lemma 3.3. *The word x is Type I if and only if $\theta(x)$ is Type I.*

Proof. Let the conjugates of x be $x_1 \prec x_2 \prec \cdots \prec x_n$. Then by Lemma 3.2 the conjugates of $\theta(x)$ are $\theta(x_1) \succ \theta(x_2) \succ \cdots \succ \theta(x_n)$. Also note that $L(x) \preceq L(y)$ implies $L(\theta(x)) \succeq L(\theta(y))$. By (3.3) x is Type I if and only if

$$x_i \prec x_j \Rightarrow L(x_i) \succeq L(x_j),$$

that is, if and only if

$$\theta(x_i) \succ \theta(x_j) \Rightarrow L(\theta(x_i)) \preceq L(\theta(x_j)),$$

that is, by (3.3), if and only if $\theta(x)$ is Type I. □

Lemma 3.4. *The word x is Type I if and only if $\phi(x)$ is Type I.*

Proof. Suppose x is Type I. Then its 2-factors come from one of the four sets in Lemma 3.1. Suppose they come from $\{ab, ac, ba, bb, ca\}$. Then the conjugates of x may be written

$$aX_1c \prec aX_2b \prec bX_3b \prec bX_4a \prec cX_5a.$$

The order is implied by x being Type I. Applying ϕ and using part (a) of Lemma 3.2 we have, using an obvious notation,

$$a\phi(X_1)ac \prec a\phi(X_2)ab \prec ab\phi(X_3)ab \prec ab\phi(X_4)a \prec ac\phi(X_5)a.$$

The full set of conjugates of $\phi(x)$ also includes $ba\phi(X_2)a$, $bab\phi(X_3)a$ and $ca\phi(x)a$. Since each word in $\phi(X_2)$ begins with a we have $ba\phi(X_2)a \prec bab\phi(X_3)a \prec ca\phi(x)a$. By inspecting the final letters of each set of conjugates we see that $\phi(x)$ is Type I. A similar argument applies if the 2-factors belong to any of the other sets in Lemma 3.1.

Now suppose that $y = \phi(x)$ is Type I. Let the lexicographically ordered conjugates of x be

$$x_1 \prec x_2 \prec \cdots \prec x_n.$$

Then by part (a) of Lemma 3.2 we have

$$\phi(x_1) \prec \phi(x_2) \prec \cdots \prec \phi(x_n)$$

and by (b) each of these is a conjugate of y. Then (3.3) tells us that

$$L(\phi(x_1)) \succeq L(\phi(x_2)) \succeq \cdots \succeq L(\phi(x_n)).$$

However, for any word u, $L(\phi(u)) = L(u)$ so

$$L(x_1) \succeq L(x_2) \succeq \cdots \succeq L(x_n).$$

This implies, by (3.3), that x is Type I. □

We now introduce the mapping ψ. Let x be a word of length n and let $i \in [1, n]$.
(a) Suppose $x[i] = a$. If $i < n$ and $x[i+1] = a$ or if $i = n$ and $x[1] = a$ then $\psi'(x[i]) = ab$; otherwise $\psi'(x[i]) = a$.
(b) Suppose $x[i] = b$. If $i < n$ and $x[i+1] \neq b$ or if $i = n$ and $x[1] \neq b$ then $\psi'(x[i]) = bb$; otherwise $\psi'(x[i]) = b$.
(c) Suppose $x[i] = c$. If $i < n$ and $x[i+1] = c$ or if $i = n$ and $x[1] = c$ then $\psi'(x[i]) = cb$; otherwise $\psi'(x[i]) = c$.

Then $\psi(x)$ is the concatenation

$$\psi'(x[1])\psi'(x[2])\dots\psi'(x[n]).$$

A more intuitive explanation of this is to say that we form $\psi(x)$ from x by inserting a b in the middle of each factor aa, cc, ba and bc and by regarding $L(x)F(x)$ as a factor. For example,

$$\psi(abbacaabac) = abbbacababbbac$$
$$\psi(aabaca) = ababbacab.$$

We will show that x is Type I if and only if $\psi(x)$ is Type I. This will require two lemmas.

Lemma 3.5. *If x is a conjugate of y then $\psi(x)$ is a conjugate of $\psi(y)$.*

Proof. This is easily checked. □

Note that $\{\psi(y) : y \in C(x)\}$ includes all conjugates of $\psi(x)$ except those with prefix ba or bc.

Lemma 3.6. *If x and y have the same length, are Type I and $x \prec y$ then $\psi(x) \prec \psi(y)$.*

Proof. If x and y have different first letters then it is easy to see that the statement holds. We therefore assume they have a non-empty common prefix. Let x and y have prefixes $u\alpha\beta$ and $u\alpha\gamma$ respectively where α, β and γ are letters with $\beta \prec \gamma$. Suppose that $\alpha = a$. We note that if a word z has prefix uaa, uab or uac then $\psi(z)$ has, respectively, prefix $vaba$, $vabb$ or vac for some word v. Since $vaba \prec vabb \prec vac$ we see that if $\alpha = a$ then $\psi(x) \prec \psi(y)$. A similar analysis shows this relation also holds when $\alpha = b$ or $\alpha = c$, and and the statement of the lemma follows. □

Lemma 3.7. *The word x is Type I if and only if $\psi(x)$ is Type I.*

Proof. Let x be a Type I word with 2-factors from the set $\{aa, ab, ac, ba, ca\}$. The conjugates of x make up sets

$$aX_1c \prec aX_2b \prec aX_3a \prec bX_4a \prec cX_5a.$$

Applying ψ to each of these sets and using Lemma 3.6 gives sets

$$aY_1c \prec aY_2bb \prec aY_3ab \prec bbY_4a \prec cY_5a, \tag{3.4}$$

where $\psi(aX_1c) = aY_1c$ et cetera. By Lemma 3.5 these are all conjugates of $\psi(x)$. To make up the full set of conjugates we include baY_2b and baY_3a. By (3.4) we have $baY_2b \prec baY_3a$, so that

$$aY_1c \prec aY_2bb \prec aY_3ab \prec baY_2b \prec baY_3a \prec bbY_4a \prec cY_5a$$

from which it follows that $\psi(x)$ is Type I.

If instead the set of 2-factors of x is a subset of $\{ab, ac, ba, bb, ca\}$ then its conjugates make up sets

$$aX_1c \prec aX_2b \prec bX_3b \prec bX_4a \prec cX_5a.$$

Applying ψ to these gives sets

$$aY_1c \prec aY_2bb \prec bbY_3b \prec bbY_4a \prec cY_5a \tag{3.5}$$

and set of conjugates baY_2b which slots in lexicographically between the second and third terms. Again $\psi(x)$ is Type I. Similar analyses apply when the set of 2-factors is one of the others in Lemma 3.1.

So far we have shown that if x is Type I then so is $\psi(x)$. We now show the converse.

Suppose that $\psi(x)$ is Type I. The definition of ψ means that $\psi(x)$ cannot contain aa or cc as 2-factors so its set of 2-factors comes from the set $\{ab, ac, ba, bb, ca\}$ or the set

$\{ac, bb, bc, ca, cb\}$. Suppose the 2-factors comes from the first of these. Then x cannot contain the factor cb as this would mean $\psi(x)$ also contains this factor which we have denied. Similarly it cannot contain bc. Neither can it contain cc as then $\psi(x)$ would contain bc. Let the set of conjugates of x be the union of sets

$$aX_1c,\ aX_2b,\ aX_3a,\ bX_4b,\ bX_5a,\ cX_6a.$$

Under ψ these give rise to the following sets of conjugates of $\psi(x)$:

$$aY_1c,\ aY_2bb,\ aY_3ab,\ bbY_4b,\ bbY_5a,\ cY_6a,$$

together with baY_2b and baY_3a. The fact that $\psi(x)$ is Type I imposes certain constraints on these sets.

We must have $aY_1c \prec aY_2bb$ and hence by Lemma 2.4 $aX_1c \prec aX_2b$ and thus $X_1 \prec X_2$. We must also have $baY_2b \prec baY_3a$ which implies $X_2 \prec X_3$. Combining these observations gives

$$X_1 \prec X_2 \prec X_3. \tag{3.6}$$

We also need $bbY_4b \prec bbY_5a$ which implies

$$X_4 \prec X_5. \tag{3.7}$$

At least one of the sets baY_3a and bbY_4b must be empty, otherwise we get a contradiction with (3.3). This means that either X_3 or X_4 is empty. The ordered set of conjugates of x is therefore

$$aX_1c \prec aX_2b \prec bX_4b \prec bX_5a \prec cX_6a$$

or

$$aX_1c \prec aX_2b \prec aX_3a \prec bX_5a \prec cX_6a.$$

By inspecting the last letters we see that x is Type I, as required. Similar arguments show that x is Type I when y has 2-factors from $\{ac, bb, bc, ca, cb\}$. □

Lemma 3.8. *Every Type I word which contains each of a, b and c has a conjugate in the range of ϕ, $\theta \circ \phi$ or ψ.*

Proof. Let y be a Type I word. We know from Lemma 3.1 that its set of 2-factors comes from one of the sets $\{aa, ab, ac, ba, ca\}$, $\{ab, ac, ba, bb, ca\}$, $\{ac, bb, bc, ca, cb\}$ and $\{ac, bc, ca, cb, cc\}$.

Suppose it comes from the first. If y does not begin with a then replace it with one of its conjugates that does. Then each occurrence of the letter b is preceded by a: we can replace such a pair with $\phi(a)$. Similarly each occurrence of c is preceded by a and the pair ac can be replaced with $\phi(c)$. The remaining occurrences of a can be replaced with $\phi(a)$ and we see that y is in the range of ϕ.

Suppose the factors of y come from the fourth set. If y does not begin with c then replace it with one of its conjugates that does. Then the factors of $\theta(y)$ come from the first set, so by the previous case there exists x such that $\phi(x) = \theta(y)$. But θ is its own inverse so $\theta \circ \phi(x) = y$ and y is in the range of $\theta \circ \phi$.

Now suppose the 2-factors of y come from the set $\{ab, ac, ba, bb, ca\}$. If y begins with ba or bc replace it with a conjugate that doesn't. Say that a factor $y[i..j]$ is a b-*run* if each of its letters equals b, but neither $y[i-1]$ nor $y[j+1]$ equals b. Construct a word x by removing a b from each b-run, except in the case where both a prefix and a suffix of y are b-runs. In this case remove a b from the prefix b-run but not from the suffix b-run. A b-run of length 1 in y will be preceded and followed by a's and correspond to a pair of a's in x. It is easy to see that $y = \psi(x)$.

If the factors come from the third set a similar argument applies but with c in the role of a. □

We have not yet shown that the words in T are primitive. The following theorem does this and will be used later to specify the possible values of i, j and k when $BWT(x) = c^i b^j a^k$. If x is in $\{a, b, c\}^*$ then the *Parikh vector* for x is the vector $p(x) = [|x|_a, |x|_b, |x|_c]$. If $p(x) = [\alpha, \beta, \gamma]$ then it is clear that

$$p(\theta(x)) = [\gamma, \beta, \alpha] \tag{3.8}$$

and

$$p(\phi(x)) = [\alpha + \beta + \gamma, \beta, \gamma]. \tag{3.9}$$

The Parikh vector for $\psi(x)$ is less obvious. Suppose that $\psi(x)$ is Type I and that its set of 2-factors comes from either $\{ab, ac, ba, bb, ca\}$ or $\{aa, ab, ac, ba, ca\}$. We write $|x|_{ab}$ for the number of occurrences of ab in x. If $L(x) = a$ and $F(x) = b$ we regard $L(x)F(x)$ as an occurrence of ab and count it in $|x|_{ab}$. We define $|x|_{aa}$ et cetera in a similar fashion. Since each occurrence of c in x is preceded and succeeded by a we have

$$|x|_{ac} = |x|_{ca} = |x|_c. \tag{3.10}$$

It is clear that $|\psi(x)|_a = \alpha$ and $|\psi(x)|_c = \gamma$. Also from the definition of ψ and (3.10),

$$\begin{aligned} |\psi(x)|_b - |x|_b &= |x|_{aa} + |x|_{cc} + |x|_{ba} + |x|_{bc} \\ &= |x|_a + |x|_c - |x|_{ca} - |x|_{ac} \\ &= |x|_a - |x|_c. \end{aligned}$$

If the 2-factors of $\psi(x)$ come from $\{ac, bb, bc, ca, cb\}$ or $\{ac, bc, ca, cb, cc\}$ then a similar equality holds with a and c interchanged. In either case we have

$$p(\psi(x)) = [\alpha, \beta + |\alpha - \gamma|, \gamma]. \tag{3.11}$$

.

Theorem 3.9. *If x is in T and $p(x) = [\alpha, \beta, \gamma]$ then* $\gcd(\alpha, \beta, \gamma) = 1$ *and* $\gcd(\alpha + \beta, \beta + \gamma) = 1$.

Proof. First suppose that x is in T_0 and recall that x is therefore a copy of a word in S. Thus one of α, β and γ equals 0 and the other two are relatively prime, so the statement of the theorem holds.

We complete the proof using induction on the sum $\alpha+\beta+\gamma$. The statement holds if the sum equals 1. Suppose it holds for all values less than n and suppose $x \in T \backslash T_0$ where $|x| = n$ and $p(x) = [\alpha, \beta, \gamma]$. We may suppose $\alpha \geq \gamma$ since otherwise we can replace x with $\theta(x)$ - clearly the statement of the theorem is true for x if and only if it is true for $\theta(x)$. Then, by the definition of T, $x = \phi(y)$ or $x = \psi(y)$ for some $y \in T$ We don't need to consider $x = \theta \circ \phi(y)$ which is impossible if $\alpha \geq \gamma$, nor $x = \theta \circ \psi(y)$ as then $x = \psi \circ \theta(y)$. By (3.9) and (3.11) it follows that $p(y)$ equals $[\alpha-\beta-\gamma, \beta, \gamma]$ or $[\alpha, \beta-(\alpha-\gamma), \gamma]$. Suppose $p(y) = [\alpha-\beta-\gamma, \beta, \gamma]$. By the induction hypothesis the components have gcd $= 1$ from which $\gcd(\alpha, \beta, \gamma) = 1$. Also by the induction hypothesis we have

$$\begin{aligned} & \gcd((\alpha-\beta-\gamma)+\beta, \beta+\gamma) = 1 \\ \Rightarrow\ & \gcd(\alpha-\gamma, \beta+\gamma) = 1 \\ \Rightarrow\ & \gcd(\alpha+\beta, \beta+\gamma) = 1. \end{aligned}$$

Thus the statement holds when when $x = \phi(y)$. The other case can be settled in the same way, and the theorem is proven by induction. □

Corollary 3.10. *The words in T are primitive.*

Proof. A word that is not primitive is a power, and each element of its Parikh vector will be divisible by the exponent of the power. By the theorem such a word cannot belong to T. □

Theorem 3.11. *The Type I words are precisely the conjugates of those in T and the powers of these conjugates.*

Proof. By Lemmas 3.3, 3.4 and 3.7 every word in T is Type I. By Lemmas 2.1 and 2.2 so are conjugates of words in T and their powers.

To prove the converse we use induction on word length to show that all primitive Type I words are conjugates of words in T. This clearly holds for words of length 1 and for words containing at most 2 distinct letters. Suppose it holds for all words of length less than n and let x' be a primitive Type I word which contains each of a, b and c and has length n. By Lemma 3.8 x' has a conjugate x for which there exists y such that either $x = \phi(y)$ or $x = \psi(y)$ or $x = \theta \circ \phi(y)$, and by the "only if" parts of Lemmas 3.3, 3.4 and 3.7 y is Type I. Clearly $|y| < n$ so by the induction hypothesis y is the conjugate of a word in T, say z. But then, by Lemmas 3.2 and 3.5 either $\phi(z) = x$, $\psi(z) = x$ or $\theta \circ \phi(z) = x$, so x is in T and x' is a conjugate of a word in T.

We have shown that every primitive Type I word is a conjugate of a word in T. To show that every non-primitive Type I word is a power of a conjugate of a word in T we apply Lemma 2.2 and argue as in the proof of Lemma 2.6. □

Theorem 3.12. *The vector $[i, j, k]$ is the Parikh vector of a Type I word if and only if* $\gcd(i, j, k) = \gcd(i+j, j+k)$.

Proof. ($\Rightarrow$) Suppose x is Type I with $p(x) = [i, j, k]$. If $x \in T$ then both greatest common divisors equal 1 by Theorem 3.9. This also applies if x is a conjugate of a word in T. Otherwise x is the dth power of such a conjugate and then both greatest common divisors equal d, as required.

($\Leftarrow$) We show that that if $\gcd(i, j, k) = \gcd(i + j, j + k)$ then there exists a Type I word x with $p(x) = [i, j, k]$. This holds if $i + j + k \leq 2$ since all words with such Parikh vectors are Type I. Suppose it holds whenever $i + j + k < n$ and consider $[i, j, k]$ with

$$\gcd(i, j, k) = \gcd(i + j, j + k) = d \tag{3.12}$$

and $i + j + k = n$. We consider 4 cases.

Case 1. If $i \geq j+k$ consider the vector $[i-j-k, j, k]$. By (3.12) $\gcd(i-j-k, j, k) = \gcd(i-k, j + k) = d$ so, providing at least one of j and k is positive, we can apply the induction hypothesis and conclude that there exists a Type I word y with $p(y) = [i-j-k, j, k]$. But then $\phi(y)$ is Type I by Lemma 3.4, and by (3.9) $p(\phi(y)) = [i, j, k]$. So $x = \phi(y)$ satisfies the statement of the theorem. If $j = k = 0$ then the Type I word $x = a^i$ is satisfactory.

Case 2. If $j + k \geq i \geq k$ consider $[i, j - i + k, k]$. By (3.12) $\gcd(i, j - i + k, k) = \gcd(j + k, j - i + 2k) = d$. Providing $i \neq k$ we can apply the induction hypothesis and conclude there exists a Type I word y with $p(y) = [i, j-i+k, k]$. Now $\psi(y)$ is Type I by Lemma 3.7 and by (3.11) $p(\psi(y)) = [i, j, k]$ so that $\psi(y)$ satisfies the requirements of the theorem. Now consider the case $i = k$. If $i = k$ and $j = 0$ the word $(ac)^i$ is Type I with Parikh vector $[i, j, k]$ as required. If $i = k = 0$ we use b^j. If $i = k > 0$ and $j > 0$ then $\gcd(i + j, j + k) = i + j > \gcd(i, j, k)$, contradicting (3.12).

Cases 3 and 4. If $k \geq i + j$ we use the vector $[k - i - j, j, i]$ and argue as in Case 1 with $\theta \circ \phi$ in the role of ϕ and if $i + j \geq k \geq i$ we use $[k, i + j - k, i]$ and argue as in Case 2 with $\theta \circ \psi$ in the role of ψ.

This completes the proof by induction. □

Note that the sets in the four cases are not disjoint. This reflects the fact that some Type I words are in the ranges of both ϕ and ψ. For example $[2, 1, 1]$ is covered by Cases 1 and 2 and $\phi(bc) = \psi(aac) = abac$. We also mention that a result equivalent to Theorem 3.9 has been obtained by Pak and Redlich [8].

4 Discussion

Some words in T are

$$baca,\ cbbca,\ cbcacca,\ bbabbaca,\ ccaccbcca,\ babacabaca,$$
$$acabacabaca,\ ababaacaaca,\ bbabbbabbaca,\ cbbbbbbbbbca,$$
$$cbcbbcbbcbca,\ cbcacbcacbcbca, ccccaccccbcccca.$$

We note that each of these has the *two palindrome property*, that is, it can be written as uv where u and v are palindromes or empty. Standard words also have this property - see Theorem 2.2.6 of [5]. We shall generalise this to larger alphabets.

We consider a word x over the alphabet $a_1 \prec a_2 \prec \cdots \prec a_s$ for which

$$BWT(x) = a_s^{m_s} a_{s-1}^{m_{s-1}} \dots a_1^{m_1} \tag{4.1}$$

for some non-negative integers m_1, m_2,..., m_s.

Let the conjugates of x be $x_1 \prec x_2 \prec \cdots \prec x_n$ where $|x| = n$ and write $F(i)$ for the first letter in x_i and $L(i)$ for the last. (Earlier we used $F(x_i)$ and $L(x_i)$.) We now encounter a notational awkwardness. Consider the lexicographically ordered conjugates of $x = aab$.

$$\begin{gathered} aab \\ aba \\ baa \end{gathered}$$

Here $F(1) = L(3)$ as they both equal a, but $F(1)$ corresponds to the first appearance of a in x and $L(3)$ to the second. If $F(i)$ and $L(j)$ correspond to the same appearance of a letter in x we write $F(i) \equiv L(j)$. Thus in the example above $F(1) \equiv L(2)$ and $F(2) \equiv L(3)$. An important property of the BW Transform is given in the following lemma.

Lemma 4.1. *If $F(i_1) \equiv L(j_1)$, $F(i_2) \equiv L(j_2)$ and $i_1 < i_2$ then $j_1 < j_2$.*

A demonstration of this result and an explanation of how it is used to invert the BW Transform are given in [2] and [6]. Equation (4.1) implies that

$$F(i) = L(n + 1 - i) \tag{4.2}$$

for $i = 1, \dots, n$. But note that we cannot replace "=" with "≡" here. For x satisfying (4.1) define ω to be the function satisfying

$$F(i) \equiv L(\omega(i)).$$

If x satisfies (4.1) then we evaluate ω as follows. Let k be the least integer such that

$$\sum_{j=1}^{k} m_j \geq i,$$

then

$$\omega(i) = i + n - \sum_{j=1}^{k-1} m_j - \sum_{j=1}^{k} m_j.$$

The straightforward derivation of this formula (using Lemma 4.1 and equation (4.1)) is omitted. We also have the following formula for ω^{-1}. Let k be the least integer such that

$$i \geq n + 1 - \sum_{j=1}^{k} m_j,$$

then

$$\omega^{-1}(i) = i - n + \sum_{j=1}^{k-1} m_j + \sum_{j=1}^{k} m_j.$$

We need the following lemma.

Lemma 4.2. *For $i \in [1, n]$ and $j \geq 1$,*

$$\omega^j(i) = n + 1 - \omega^{-j}(n + 1 - i). \tag{4.3}$$

Proof. Using the formulae above we obtain

$$\omega(i) = n + 1 - \omega^{-1}(n + 1 - i). \tag{4.4}$$

(In obtaining this we find the value of k in each formula to be the same. One can also see (4.4) immediately by recognising the symmetry imposed by Lemma 4.1 and equation (4.1)). From (4.4) we prove (4.3) by induction. Suppose it holds for $j = k$. Then

$$\begin{aligned} \omega^{k+1}(i) &= \omega(\omega^k(i)) \\ &= n + 1 - \omega^{-1}(n + 1 - \omega^k(i)) \\ &= n + 1 - \omega^{-1}(\omega^{-k}(n + 1 - i)) \\ &= n + 1 - \omega^{-(k+1)}(n + 1 - i), \end{aligned}$$

as required. □

We can use ω to obtain x_i. The first letter in x_i is $F(i)$ which equals $L(\omega(i))$, and $L(\omega(i))$ is followed by $F(\omega(i))$. Continuing in this way we find

$$x_i = F(i)F(\omega(i))F(\omega^2(i)) \ldots F(\omega^{n-1}(i)). \tag{4.5}$$

Similarly, the reverse of x_i is

$$x_i^R = L(i)L(\omega^{-1}(i)) \ldots L(\omega^{-(n-1)}(i)). \tag{4.6}$$

We now obtain two important results about words satisfying (4.1).

Theorem 4.3. *If x satisfies (4.1) and has lexicographically ordered conjugates $x_1, x_2, \ldots,$ x_n then, for $i = 1, \ldots, n$,*

$$x_i = x_{n+1-i}^R.$$

Proof. Using (4.6), Lemma 4.2, (4.2) and (4.5) we have

$$\begin{aligned} x_{n+1-i}^R &= L(n + 1 - i)L(\omega^{-1}(n + 1 - i)) \ldots L(\omega^{-(n-1)}(n + 1 - i)) \\ &= L(n + 1 - i)L(n + 1 - \omega(i)) \ldots L(n + 1 - \omega^{n-1}(i)) \\ &= F(i)F(\omega(i)) \ldots F(\omega^{n-1}(i)) \\ &= x_i. \end{aligned}$$

□

Corollary 4.4. *Under the conditions of the theorem each conjugate of x has the two palindrome property.*

Proof. Consider x_i. If $i = n + 1 - i$ then by the theorem $x_i = x_i^R$ so x_i is a palindrome and therefore has the two palindrome property. Now suppose $i \neq n+1-i$ so that x_i and x_{n+1-i} are different conjugates of x. Then $x_i = uv$ and $x_{n+1-i} = vu$ for some u and v. By the theorem

$$uv = (vu)^R = u^R v^R$$

so that u and v are palindromes, and x_i has the two palindrome property. □

As noted above the binary alphabet case of Corollary 4.4 is known. The binary case of Theorem 4.3 has been obtained by Restivo and Sciortino [10] and reported at the conference WORDS 2005 in Montreal, and at the Workshop on Fibonacci Words, Turku, September 2006. They found various other interesting symmetries that apply in this case but do not hold with larger alphabets.

The *complexity* of a word x is a function $c(n)$ equalling the number of distinct factors of length n appearing x. An important property of standard words is that their complexity satisfies $c(n) \leq n+1$ for all n. Indeed standard words are factors of Sturmian words and for Sturmian words $c(n) = n+1$ for all n. We will show that for Type I words $c(n) \leq 2n+1$ for all n. In fact we prove a stronger result in Theorem 4.5 below. But first some notation. A finite word w is a *left special factor* of a word x if there exists more than one distinct letter α such that αw is a factor of x. The number of such α is the *left degree* of w, written $\lambda(w)$. Thus w is left special if and only if $\lambda(w)$ is greater than 1. From these definitions we have, for a word x,

$$c(n+1) = \sum_{|w|=n} \lambda(w), \tag{4.7}$$

where the sum is over all distinct length n factors of x. Right special factors and right degrees are defined in a similar way.

Theorem 4.5. *If x is a word on the alphabet $\{a_1, a_2, \ldots, a_k\}$ for which $BWT(x) = a_k^{m_k} \ldots a_1^{m_1}$ for some integers $m_1, \ldots, m_k$ then*

$$c(n) \leq (k-1)n + 1$$

for all n.

Proof. Consider a left special factor w of x which has length n and degree t, and consider those conjugates of x which have prefix w. In lexicographic order these have the form

$$\begin{array}{l} wv_1a_t \\ wv_2a_t \\ \vdots \\ wv_{k_1}a_t \\ wv_{k_1+1}a_{t-1} \\ \vdots \\ wv_{k_2}a_{t-1} \\ \vdots \\ wv_{k_t}a_1 \end{array}$$

Note that the last column here has changed value $t-1$ times. But this column is part of $BWT(x)$, which contains at most $k-1$ such changes. Each w will contribute $\lambda(u)-1$ such changes, so we have

$$\sum_{|u|=n}(\lambda(u)-1) \leq k-1$$

where the sum is over all distinct length n factors u of x. Noting that $\sum_{|u|=n} 1 = c(n)$ and applying (4.7) gives

$$c(n+1) \leq c(n)+k-1.$$

Since $c(1) = k$ we obtain, by induction, that $c(n) \leq (k-1)n+1$ for all n. □

On setting $k = 3$ we obtain the following corollary.

Corollary 4.6. *If x is a Type I word then its complexity satisfies $c(n) \leq 2n+1$ for all n.*

Note that we haven't assumed any ordering on $a_1, a_2, \ldots, a_k$ in the theorem, so the result applies also, for example, to a word x for which $BWT(x) = c^ia^jb^k$. Such words do exist, for example *abcbc*. So do words x for which $BWT(x) = b^ic^ja^k$, for example *bacbabba*. Other orderings of a, b and c are not possible. We call such words Types II and III respectively. We haven't made a thorough investigation of these, but note that (a) the morphism θ transforms a Type II word into a Type III word, and vice versa, and (b) these words do not, in general, have the two palindrome property.

We mention that a different set of words satisfying $c(n) \leq 2n+1$ appears in [1] and is discussed in [3] where they are called 3-standard words. Like ours they have some properties in common with standard words. For example standard words are connected with the Fine-Wilf Periodicity Lemma [4] which states that if a word has periods p and q and length at least $p+q-\gcd(p,q)$ then it has period $\gcd(p,q)$. If a word w has periods p and q with $\gcd(p,q) = 1$, length $p+q-2$ and does not have period 1 then wab or wba is a standard word. For example *abaaba* has periods 3 and 5 and length $3+5-2$. Both *abaabaab* and *abaababa* are standard words. In [3] a three period version of the Periodicity Lemma is proved and the words which are critical for it are the 3-standard words. These have the two palindrome property but not the property described in Theorem 4.3.

There are several directions in which this work might be extended. First, can we find characterisations of words on larger alphabets which have simple Burrows-Wheeler Transforms and can we characterise the Type II and Type III words mentioned above? For larger alphabets can we find results like Theorem 3.12? For binary alphabets standard words can be extended to Sturmian words which are infinite with complexity $c(n) = n+1$ for all n. Can our Type I words be used to construct infinite words with complexity $2n+1$ for all n?

Acknowledgement. We thank Professor Antonio Restivo for useful discussions and for kindly alerting us to reference [8]. He has also informed us that Giovanna Rosone has independently proved Theorem 4.3 and Corollary 4.4.

References

[1] P. Arnoux and G. Rauzy, *Représentation géométrique de suites de complexité 2n+1*, Bull. Soc. Math. France 119 (1991) 199-215.

[2] M. Burrows and David J. Wheeler, *A block-sorting lossless data compression algorithm*, HP Lab Technical Report, 1994, available on `http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.html`

[3] M.G. Castelli, F. Mignosi and A. Restivo, *Fine and Wilf's theorem for three periods and a generalisation of Sturmian words*, Th. Comput. Sci. 218(1999), 3-12.

[4] N.J. Fine and H.S. Wilf, *Uniqueness theorem for periodic functions*, Proc. Amer. Math. Soc., 16(1965) 109-114.

[5] M. Lothaire, *Algebraic Combinatorics on Words*, Encyclopedia of Mathematics and its Applications 90, Cambridge, 2002.

[6] S. Mantaci, A. Restivo and M. Sciortino, *Burrows-Wheeler Transform and Sturmian words* Information Proc. Letters, 86(2003) 241-246.

[7] G. Manzini and T. Gagie, Move-to-front, distance coding and inversion frequencies revisited, In B. Ma and K. Zhang, editors, *Combinatorial Pattern Matching*, number 4580 in Lecture Notes in Computer Science, pages 71–82. Springer-Verlag, Berlin, 2007.

[8] I. Pak and A. Redlich, *Long cycles in abc-permutations*, preprint, 2007, available on `http://www-math.mit.edu/~pak/research.html`

[9] G. Navarro and V. Mäkinen, Compressed Full Text Indexing, *ACM Computing Surveys*, 39(1):1–61, 2007.

[10] A. Restivo, personal communication.