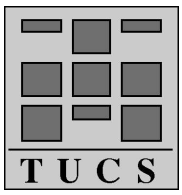
Periodicity and Unbordered Words

Tero Harju
Dirk Nowotka
Turku Centre for Computer Science, TUCS,
Department of Mathematics, University of Turku



Turku Centre for Computer Science
TUCS Technical Report No 523
April 2003
ISBN 952-12-1154-7
ISSN 1239-1891

Abstract

The relationship between the length of a word and the maximum length of its unbordered factors is investigated in this paper.

Consider a finite word w of length n. Let $\mu(w)$ denote the maximum length of its unbordered factors, and let $\partial(w)$ denote the period of w. Clearly, $\mu(w) \leq \partial(w)$.

We establish that $\mu(w) = \partial(w)$, if w has an unbordered prefix of length $\mu(w)$ and $n \geq 2\mu(w) - 1$. This bound is tight and solves a 21 year old conjecture by Duval. It follows from this result that, in general, $n \geq 3\mu(w) - 2$ implies $\mu(w) = \partial(w)$ which gives an improved bound for the question asked by Ehrenfeucht and Silberger in 1979.

Keywords: combinatorics on words, periodicity, unbordered factors, Duval's conjecture

TUCS Laboratory
Discrete Mathematics for Information Technology

1 Introduction

Periodicity and borderedness are two properties of words—the most basic data structure—which are investigated in this paper. These concepts are so foundational that they play a rôle (explicitely or implicitely) in virtually every area of computer science. Just a few of those areas are string searching algorithms [15, 3, 8], data compression [23, 7], and codes [2], which are classical examples, but also computational biology, e.g., sequence assembly [19] or superstrings [4], and serial data communications systems [5] are areas among others where periodicity and borderedness of words (sequences) are important concepts. It is well known that these two word properties do not exist independently from each other. However, it is somewhat surprising that no clear relation has been established so far, despite the fact that this basic question has been around for more than 20 years.

Let us consider a finite word (a sequence of letters) w. We denote the length of w by $|w|$ and call a subsequence of consecutive letters of a word *factor*. The period of w, denoted by $\partial(w)$, is the smallest positive integer p such that the i-th letter equals the $(i+p)$-th letter for all $1 \leq i \leq |w| - p$. Let $\mu(w)$ denote the length of the longest unbordered factor of w. A word is bordered, if it has a proper prefix that is also a suffix, where we call a prefix proper, if it is neither empty nor contains the entire word. For the investigation of the relationship between $|w|$ and the maximality of $\mu(w)$, that is, $\mu(w) = \partial(w)$, we consider the special case where the longest unbordered prefix of a word is of the maximum length, that is, no unbordered factor is longer than that prefix. Let w be an unbordered word. Then a word wu is a *Duval extension* (of w), if every unbordered factor of wu has at most length $|w|$, that is, $\mu(wu) = |w|$. We call wu *trivial* Duval extension, if $\partial(wu) = |w|$. For example, let $w = abaabb$ and $u = aaba$. Then $wu = abaabbaaba$ is a nontrivial Duval extension of w since (i) w is unbordered, (ii) all factors of wu longer than w are bordered, that is, $|w| = \mu(wu) = 6$, and (iii) the period of wu is 7, and hence, $\partial(wu) > |w|$. Note, that this example satisfies $|u| = |w| - 2$.

In 1979 Ehrenfeucht and Silberger initiated a line of research [11, 1, 10] exploring the relationship between the length of a word w and $\mu(w)$. In 1982 these efforts culminated in Duval's result: If $|w| \geq 4\mu(w) - 6$ then $\partial(w) = \mu(w)$. However, it was conjectured in [1] that $|w| \geq 3\mu(w)$ implies $\partial(w) = \mu(w)$ which follows if Duval's conjecture [10] holds true.

Conjecture 1. *Let wu be a nontrivial Duval extension of w. Then $|u| < |w|$.*

After that, no progress was recorded, to the best of our knowledge, for 20 years. However, the topic remained popular, see for example Chapter 8

in [17]. The most recent results are by Mignosi and Zamboni [20] and the authors of this article [13]. However, not Duval's conjecture but rather its opposite is investigated in those papers, that is: Which words admit only trivial Duval extensions? It is shown in [20] that unbordered, finite factors of Sturmian words allow only trivial Duval extensions, with other words, if an unbordered, finite factor of a Sturmian word of length $\mu(w)$ is a prefix of w, then $\partial(w) = \mu(w)$. Sturmian words are binary infinite words of minimal complexity; see [21] and Chapter 2 in [17]. That result was improved in [13] by showing that Lyndon words [18] allow only trivial Duval extensions and the fact that every unbordered, finite factor of a Sturmian word is a Lyndon word.

The main result in this paper is an improved version of Conjecture 1.

Theorem 2. *Let wu be a Duval nontrivial extension of w. Then $|u| < |w|-1$.*

The example mentioned above shows that this bound on the length of a nontrivial Duval extension is tight. Theorem 2 implies the truth of Duval's conjecture, as well as, the following corollary (for any word w).

Corollary 3. *If $|w| \geq 3\mu(w) - 2$, then $\partial(w) = \mu(w)$.*

This corollary confirms the conjecture by Assous and Pouzet in [1] about a question asked by Ehrenfeucht and Silberger in [11].

Our main result, Theorem 2, is presented in Section 4, which uses the notations introduced in Section 2 and preliminary results from Section 3. We conclude with Section 5.

2 Notations

In this section we introduce the notations of this paper. We refer to [16, 17] for more basic and general definitions.

We consider a finite alphabet A of letters. Let A^* denote the monoid of all finite words over A including the empty word, denoted by ε. Let $w = w_{(1)}w_{(2)}\cdots w_{(n)}$ where $w_{(i)}$ is a letter, for every $1 \leq i \leq n$. We denote the length n of w by $|w|$. An integer $1 \leq p \leq n$ is a *period* of w, if $w_{(i)} = w_{(i+p)}$ for all $1 \leq i \leq n-p$. The smallest period of w is called the *minimum period* (or simply, the period) of w, denoted by $\partial(w)$. A nonempty word u is called a *border* of a word w, if $w = uv = v'u$ for some suitable words v and v'. We call w *bordered*, if it has a border that is shorter than w, otherwise w is called *unbordered*. Note, that every bordered word w has a minimum border u such that $w = uvu$, where u is unbordered. Let $\mu(w)$ denote the maximum length of unbordered factors of w. Suppose $w = uv$, then u is called a *prefix* of w,

denoted by $u \leq w$, and v is called a *suffix* of w, denoted by $v \preccurlyeq w$. Let $u, v \neq \varepsilon$. Then we say that u *overlaps* v *from the left* or *from the right*, if there is a word w such that $|w| < |u| + |v|$, and $u \leq w$ and $v \preccurlyeq w$, or $v \leq w$ and $u \preccurlyeq w$, respectively. We say that u *overlaps* (intersects) with v, if either v is a factor of u or u is a factor of v or u overlaps v from the left or right.

Let us consider the following examples. Let $A = \{a, b\}$ and $u, v, w \in A^*$ such that $u = abaa$ and $v = baaba$ and $w = abaaba$. Then $|w| = 6$, and 3, 5, and 6 are periods of w, and $\partial(w) = 3$. We have that a is the shortest border of u and w, whereas ba is the shortest border of v. We have $\mu(w) = 3$. We also have that u and v overlap since $u \leq w$ and $v \preccurlyeq w$ and $|w| < |u| + |v|$.

We continue with some more notations. Let w and u be nonempty words where w is also unbordered. We call wu a *Duval extension* of w, if every factor of wu longer than $|w|$ is bordered, that is, $\mu(wu) = |w|$. A Duval extension wu of w is called *trivial*, if $\partial(wu) = \mu(wu) = |w|$. A nontrivial Duval extension wu of w is called *minimal*, if u is of minimal length, that is, $u = u'a$ and $w = u'bw'$ where $a, b \in A$ and $a \neq b$.

Example 4. *Let $w = abaabbabaababb$ and $u = aaba$. Then*

$$w.u = abaabbabaababb.aaba$$

(for the sake of readability, we use a dot to mark where w ends) is a nontrivial Duval extension of w of length $|wu| = 18$, where $\mu(wu) = |w| = 14$ and $\partial(wu) = 15$. However, wu is not a minimal Duval extension, whereas

$$w.u' = abaabbabaababb.aa$$

is minimal, with $u' = aa \leq u$. Note, that wu is not the longest nontrivial Duval extension of w since

$$w.v = abaabbabaababb.abaaba$$

is longer, with $v = abaaba$ and $|wv| = 20$ and $\partial(wv) = 17$. One can check that wv is a nontrivial Duval extension of w of maximum length, and at the same time wv is also a minimal Duval extension of w.

Let an integer p with $1 \leq p < |w|$ be called *point* in w. Intuitively, a point p denotes the place between $w_{(p)}$ and $w_{(p+1)}$ in w. A nonempty word u is called a *repetition word* at point p if $w = xy$ with $|x| = p$ and there exist x' and y' such that $u \preccurlyeq x'x$ and $u \leq yy'$. For a point p in w, let

$$\partial(w, p) = \min\big\{|u| \;\big|\; u \text{ is a repetition word at } p\big\}$$

denote the *local period* at point p in w. Note, that the repetition word of length $\partial(w,p)$ at point p is necessarily unbordered and $\partial(w,p) \leq \partial(w)$. A factorization $w = uv$, with $u, v \neq \varepsilon$ and $|u| = p$, is called *critical*, if $\partial(w,p) = \partial(w)$, and, if this holds, then p is called *critical point*.

Example 5. *The word*

$$w = ab.aa.b$$

has the period $\partial(w) = 3$ *and two critical points,* 2 *and* 4*, marked by dots. The shortest repetition words at the critical points are* aab *and* baa*, respectively. Note, that the shortest repetition words at the remaining points* 1 *and* 3 *are* ba *and* a*, respectively.*

3 Preliminary Results

We state some auxiliary and well-known results about repetitions and borders in this section which will be used to prove Theorem 2, in Section 4. The proofs of these auxiliary results are straightforward and not given in this extended abstract. Results taken from the literature are referenced to.

Lemma 6. *Let* $zf = gzh$ *where* $f, g \neq \varepsilon$*. Let* az' *be the maximum unbordered prefix of* az*. If* az *does not occur in* zf*, then* agz' *is unbordered.*

Proof. Assume agz' is bordered, and let y be its shortest border. In particular, y is unbordered. If $|z'| \geq |y|$ then y is a border of az' which is a contradiction. If $|az'| = |y|$ or $|az| < |y|$ then az occurs in zf which is again a contradiction. If $|az'| < |y| \leq |az|$ then az' is not maximum since y is unbordered; a contradiction. □

The proof of the following lemma is easy.

Lemma 7. *Let* w *be an unbordered word and* $u \leq w$ *and* $v \preccurlyeq w$*. Then* uw *and* wv *are unbordered.*

The critical factorization theorem is one of the main results about periodicity of words. A weak version of it was first conjectured by Schützenberger [22] and proved by Césari and Vincent [6]. It was developed into its current form by Duval [9]. We refer to [12] for a short proof of the CFT.

Theorem 8 (CFT). *Every word* w*, with* $|w| \geq 2$*, has at least one critical factorization* $w = uv$*, with* $u, v \neq \varepsilon$ *and* $|u| < \partial(w)$*, i.e.,* $\partial(w, |u|) = \partial(w)$*.*

We have the following two lemmas about properties of critical factorizations.

Lemma 9. *Let $w = uv$ be unbordered and $|u|$ be a critical point of w. Then u and v do not overlap.*

Proof. Note, that $\partial(w, |u|) = \partial(w) = |w|$ since w is unbordered. Let $|u| \leq |v|$ without restriction of generality. Assume that u and v overlap. If $u = u's$ and $v = sv'$, then $\partial(w, |u|) \leq |s| < |w|$. On the other hand, if $u = su'$ and $v = v's$, then w is bordered with s. Finally, if $v = sut$ then $\partial(w, |u|) \leq |su| < |w|$. □

The next result follows directly from Lemma 9.

Lemma 10. *Let u_0u_1 be unbordered and $|u_0|$ be a critical point of u_0u_1. Then for any word x, we have u_ixu_{i+1}, where the indices are modulo 2, is either unbordered or has a minimum border g such that $|g| \geq |u_0| + |u_1|$.*

The next theorem states a basic fact about minimal Duval extensions. See [14] for a proof of it.

Theorem 11. *Let wu be a minimal Duval extension of w. Then u occurs in w.*

The following Lemmas 12, 13 and 14 and Corollary 3 are given in [10]. Let $a_0, a_1 \in A$, with $a_0 \neq a_1$, and $t_0 \in A^*$. Let the sequences (a_i), (s_i), (s'_i), (s''_i), and (t_i), for $i \geq 1$, be defined by

- $a_i = a_{i \pmod 2}$, that is, $a_i = a_0$ or $a_i = a_1$, if i is even or odd, respectively,
- s_i such that a_is_i is the shortest border of a_it_{i-1},
- s'_i such that $a_{i+1}s'_i$ is the longest unbordered prefix of $a_{i+1}s_i$,
- s''_i such that $s'_is''_i = s_i$,
- t_i such that $t_is''_i = t_{i-1}$.

For any parameters of the above definition, the following holds.

Lemma 12. *For any a_0, a_1, and t_0 there exists an $m \geq 1$ such that*

$$|s_1| < \cdots < |s_m| = |t_{m-1}| \leq \cdots \leq |t_0|$$

and $s_m = t_{m-1}$ and $|t_0| \leq |s_m| + |s_{m-1}|$.

Lemma 13. *Let $z \leq t_0$ such that a_0z and a_1z do not occur in t_0. Let a_0z_0 and a_1z_1 be the longest unbordered prefixes of a_0z and a_1z, respectively. Then*

1. *if $m = 1$ then a_0t_0 is unbordered,*

2. *if $m > 1$ is odd, then $a_1 s_m$ is unbordered and $|t_0| \leq |s_m| + |z_0|$,*

3. *if $m > 1$ is even, then $a_0 s_m$ is unbordered and $|t_0| \leq |s_m| + |z_1|$.*

Lemma 14. *Let v be an unbordered factor of w of length $\mu(w)$. If v occurs twice in w, then $\mu(w) = \partial(w)$.*

Corollary 15. *Let wu be a Duval extension of w. If w occurs twice in wu, then wu is a trivial Duval extension.*

4 Main Result

The next theorem proves Duval's conjecture.

Theorem 2. *Let wu be a nontrivial Duval extension of w. Then $|u| < |w|-1$.*

Proof. Recall that every factor of wu which is longer than $|w|$ is bordered since wu is a Duval extension of w. Let z be the longest suffix of w that occurs twice in zu.

If $z = \varepsilon$ then $a \preccurlyeq w$ and $u = b^j$, where $a, b \in A$ and $a \neq b$ and $j \geq 1$, but now $|u| < |w|$ since ab^j is unbordered. Moreover, $w = b^k a w' a$ with $k < j$, otherwise wu is a trivial Duval extension, and either $aw'ab^j$ is bordered, in this case it follows $j \leq |w'|$, or $aw'ab^j$ is unbordered. In both cases it follows $|u| < |w| - 1$.

So, assume $z \neq \varepsilon$. We have $z \neq w$ since wu is otherwise trivial by Corollary 3. Let $a, b \in A$ be such that

$$w = w'az \qquad \text{and} \qquad u = u'bzr$$

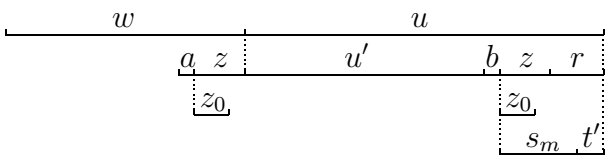
and z occurs in zr only once, that is, bz matches the rightmost occurrence of z in u. Note, that bz does not overlap az from the right, by Lemma 7, and therefore u' exists, although it might be empty. Naturally, $a \neq b$ by the maximality of z, and $w' \neq \varepsilon$, otherwise $azu'bz \leq wu$ has either no border or w is bordered (if $azu'bz$ has a border not longer than z) or az occurs in zu (if $azu'bz$ has a border longer than z); a contradiction in any case.

Let az_0 and bz_1 denote the longest unbordered prefix of az and bz, respectively. Let $a_0 = a$ and $a_1 = b$ and $t_0 = zr$ and the integer m be defined as in Lemma 13. We have then a word s_m, with its properties defined by Lemma 13, such that

$$t_0 = s_m t' \ .$$

Consider $azu'bz_0$. We have that az and $azu'bz_0$ are both prefixes of $a_0 zu$, and bz_0 is a suffix of $azu'bz_0$ and az does not occur in $zu'bz_0$. It follows from Lemma 6 that $azu'bz_0$ is unbordered, and hence,

$$|azu'bz_0| \leq |w| \ . \tag{1}$$



Case: Suppose that m is even. Then we have $2 \leq m$ and as_m $(= a_m s_m)$ is unbordered and $|t_0| \leq |s_m| + |z_1|$ by Lemma 13.

Suppose $|t_0| = |s_m| + |z_1|$ and $z_1 = z$. Then $|s_{m-1}| = |z|$ by Lemma 12. Note, that $s_i \leq t_{i-1} \leq t_0$ for all $1 \leq i \leq m$, and hence, it follows that $s_i \leq z$ for all $1 \leq i < m$. In particular, $s_{m-1} = z$. We have that bz $(= a_1 s_{m-1})$ is a border of bt_{m-2} $(= a_1 t_{m-2})$. But now, bz occurs in t_0, and hence, in u, since $t_i \leq t_0$, for all $0 \leq i < m$, which is a contradiction.

So, assume that $|t_0| < |s_m| + |z_1|$ or $|z_1| < |z|$. Suppose $|s_m| \leq |z_0|$. Then $|azu'bz_0| \leq |w|$ and

$$\begin{aligned} |u| &= |azu| - |z| - 1 \\ &= |azu'bz_0| - |z_0| + |t_0| - |z| - 1 \\ &< |azu'bz_0| - |z_0| + |s_m| + |z_1| - |z| - 1 \\ &\leq |w| + |z_1| - |z| - 1 \\ &\leq |w| - 1 \end{aligned}$$

if $|t_0| < |s_m| + |z_1|$, or

$$\begin{aligned} |u| &= |azu| - |z| - 1 \\ &= |azu'bz_0| - |z_0| + |t_0| - |z| - 1 \\ &\leq |azu'bz_0| - |z_0| + |s_m| + |z_1| - |z| - 1 \\ &\leq |w| + |z_1| - |z| - 1 \\ &< |w| - 1 \end{aligned}$$

if $|z_1| < |z|$. We have $|u| < |w| - 1$ in both cases.

Let then $|s_m| > |z_0|$. We have that as_m is unbordered, and since az_0 is the longest unbordered prefix of az, we have $az \leq as_m$, and hence, $|z| \leq |s_m|$. Now, $azu'bs_m$ is unbordered otherwise its shortest border is longer than az, since no prefix of az is a suffix of as_m, and az occurs in u; a contradiction. So, $|azu'bs_m| \leq |w|$ and $|u| < |w| - 1$, since either $|z_1| \leq |z|$ or $|t_0| < |s_m| + |z_1|$.

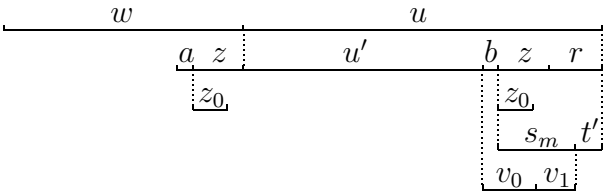
Case: Suppose that m is odd. Then bs_m $(= a_m s_m)$ is unbordered and $|t_0| \leq |s_m| + |z_0|$; see Lemma 13. Surely $s_m \neq \varepsilon$.

If $|s_m| < |z|$, then $|u| < |w| - 1$ since

$$|u| = |azu'bz_0| - |bz_0| + |bt_0| - |az|$$

and $|azu'bz_0| \leq |w|$, by (1), and $|t_0| \leq |s_m| + |z_0|$.

Assume thus that $|s_m| \geq |z|$, and hence, also $z \leq s_m$. Since $s_m \neq \varepsilon$, we have $|bs_m| \geq 2$, and therefore, by the critical factorization theorem, there exists a critical point p in bs_m such that $bs_m = v_0v_1$, where $|v_0| = p$.



In particular,

$$bz \leq v_0v_1 \ . \tag{2}$$

Note, that if $s_m = z$ then $|z_0| < |z|$ since $b \preccurlyeq z_0$ and bs_m does not end with b because it is unbordered. We have therefore in all cases

$$|z_0| < |v_0v_1| - 1 \ . \tag{3}$$

Let

$$u = u_0'v_0v_1u_1$$

be such that v_0v_1 does not occur in u_0'. Note, that v_0v_1 does not overlap with itself since it is unbordered, and v_0 and v_1 do not overlap by Lemma 9. Consider the prefix $wu_0'bz$ of wu which is bordered and has a shortest border g longer than z, and hence, $bz \preccurlyeq g$, otherwise w is bordered since $z \preccurlyeq w$. Moreover, $g \leq w$, for otherwise az would occur in u, and hence, bz occurs in w. Let

$$w = w_0bzw_1$$

such that bz occurs in w_0bz only once, that is, we consider the leftmost occurrence of bz in w. Note, that

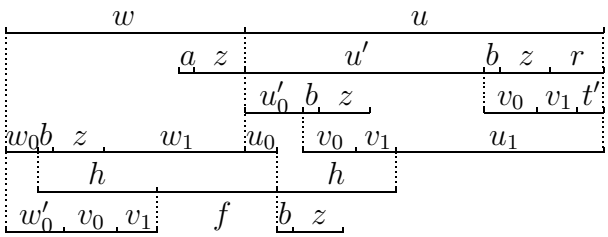
$$|w_0bz| \leq |g| \leq |u_0'bz| \tag{4}$$

where the first inequality comes from the definition of w_0 above and the second inequality from the fact that $|u_0'bz| < |g|$ implies that w is bordered. Let

$$f = bzw_1u_0'v_0v_1 \ .$$

If f is unbordered, then $|f| \leq |w|$, and hence, $|u_0'v_0v_1| \leq |w_0|$. Now, we have $|u_0'| < |w_0|$ which contradicts (4).

Therefore, f is bordered. Let h be its shortest border.



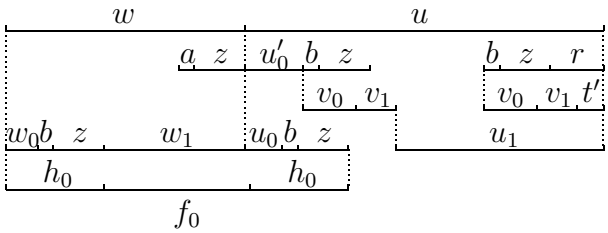
Surely, $|bz| < |h|$ otherwise v_0v_1 is bordered by (2). So, $bz \leq h$. Moreover, $|v_0v_1| \leq |h|$ otherwise bz occurs in s_m contradicting our assumption that bzr marks the rightmost occurrence of bz in u. So, $v_0v_1 \preccurlyeq h$, and v_0v_1 occurs in w since $w_0h \leq w$ by (4). Let

$$w_0bzv' = w_0h = w'_0v_0v_1 \ .$$

Note, that v_0v_1 does not occur in w'_0 otherwise it occurs in u'_0 contradicting our assumption on u'_0. Moreover, we have $h = bzv' \preccurlyeq u'_0v_0v_1$. Let $u'_0v_0v_1 = u_0h$. Consider

$$f_0 = wu_0bz$$

which has a shortest border h_0.



Surely, $bz \preccurlyeq h_0$ otherwise w is bordered with a suffix of z. Moreover, $|w_0bz| \leq |h_0| \leq |u_0bz|$ since bz does not occur in w_0 and w is unbordered. From that and $w_0h = w'_0v_0v_1$ and $u_0h = u'_0v_0v_1$ follows now $|w'_0| \leq |u'_0|$ and

$$u'_0v_0v_1 = u_0bzv' \text{ and } w_0 \text{ occurs in } u_0. \tag{5}$$

Let now

$$w = w'_0v_0v_1w'_i \cdots v_0v_1w'_2v_0v_1w'_1v_0v_1w_2$$

for some word w_2 that does not contain v_0v_1, and

$$u = u'_0v_0v_1u'_j \cdots v_0v_1u'_2v_0v_1u'_1v_0v_1t'$$

such that v_0v_1 does not occur in w'_k, for all $0 \leq k \leq i$, or v'_ℓ, for all $0 \leq \ell \leq j$. Note, that these factorizations of w and u are unique, and, moreover, $w_2 \neq \varepsilon$. (Indeed, if $w_2 = \varepsilon$ then $v_0v_1 \preccurlyeq w$ and $az \preccurlyeq v_0v_1$, and az would occur in u; a contradiction.)

We claim that either $i = j$ and $w'_k = u'_k$, for all $1 \leq k \leq i$ or $|u| < |w| - 1$. Assume $k = 1$. We show that $w'_1 = u'_1$. Consider

$$f_1 = v_1 w'_1 v_0 v_1 w_2 u'_0 v_0 v_1 u'_j \cdots v_0 v_1 u'_1 v_0 \ .$$

If f_1 is unbordered, then $|u| < |w| - 1$ since $|f_1| \leq |w|$ and

$$|u| = |f_1| - |v_1 w'_1 v_0 v_1 w_2| + |v_1 t'|$$

and $|t'| \leq |z_0| \leq |z| < |bz| \leq |v_0 v_1|$ and $w_2 \neq \varepsilon$. Assume then that f_1 is bordered, and let h_1 be its shortest border. Clearly, $h_1 = v_1 g_1 v_0$ for some g_1 (possibly $g_1 = \varepsilon$) since v_0 and v_1 do not overlap. We show that $h_1 \leq v_1 w'_1 v_0$. Indeed, otherwise either

1. az occurs in u, in case $v_1 w'_1 v_0 v_1 w_2 \leq h_1$, a contradiction to our assumption on az, or

2. v_0 and v_1 overlap, in case $|v_0| \leq |z|$ and

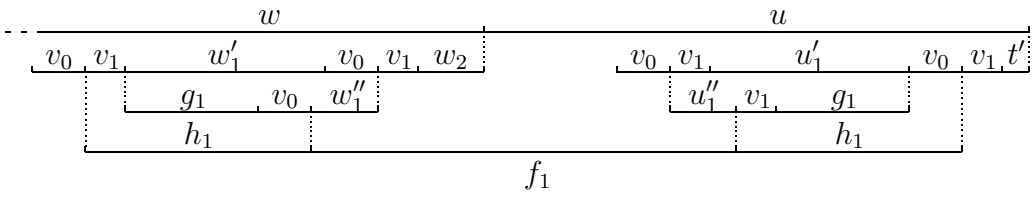
   $$|v_1 w'_1 v_0 v_1 w_2| - |az| + |v_0| < |h_1| < |v_1 w'_1 v_0 v_1 w_2|$$

   and then v_0 occurs in z, contradicting Lemma 9, or

3. $|u| < |w| - 1$, in case $v_0 w_3 \preccurlyeq w_2$ and $|az| \leq |v_0 w_3|$, then $v_0 w_3 u' v_0 v_1$ is unbordered and the result follows from $|t'| < |v_0 w_3| - 1$, since $|az| \neq |v_0 w_3|$ for v_0 does not begin with a.

Moreover, $h_1 \preccurlyeq v_1 u'_1 v_0$ since $v_0 v_1$ does not occur in $v_1 w'_1 v_0$. So, let

$$w'_1 v_0 = g_1 v_0 w''_1 \qquad \text{and} \qquad v_1 u'_1 = u''_1 v_1 g_1 \ . \tag{6}$$



Consider,

$$f_2 = v_0 w''_1 v_1 w_2 u'_0 v_0 v_1 u'_j \cdots v_0 v_1 u'_1 v_0 v_1 \ .$$

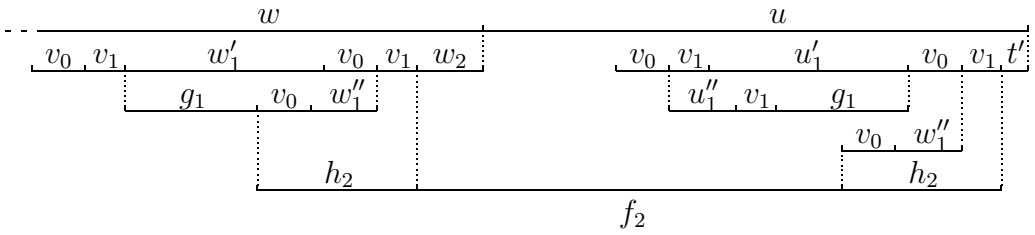
If f_2 is unbordered, then $|u| < |w| - 1$ since $|f_2| \leq |w|$ and

$$|u| = |f_2| - |v_0 w''_1 v_1 w_2| + |t'|$$

and $|t'| \leq |z_0| \leq |z| < |bz| \leq |v_0 v_1|$ and $w_2 \neq \varepsilon$. Assume then that f_2 is bordered, and let h_2 be its shortest border. Since v_0 and v_1 do not overlap,

$v_0v_1 \preccurlyeq h_2$. Also $h_2 \leq v_0w_1''v_1$ since v_0v_1 does not occur in w_2 (and v_0 and v_1 do not overlap) and az does not occur in h_2 (and so h_2 does not stretch beyond w). We have $v_0w_1''v_1 \leq h_2$ since v_0v_1 does not occur in $v_0w_1''v_1$ unless $w_1'' = \varepsilon$. Hence, we have $h_2 = v_0w_1''v_1$ and

$$w_1'v_0v_1 = g_1h_2 \qquad \text{and} \qquad h_2 \preccurlyeq u_1'v_0v_1 \ . \tag{7}$$



Consider,

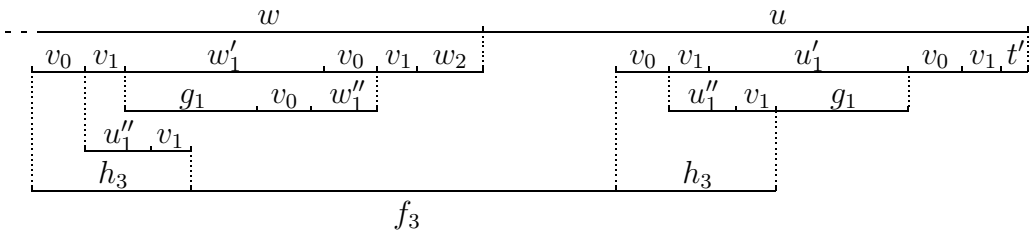
$$f_3 = v_0v_1w_1'v_0v_1w_2u_0'v_0v_1u_j' \cdots v_0v_1u_2'v_0u_1''v_1 \ .$$

If f_3 is unbordered, then $|u| < |w| - 1$ since $|f_3| \leq |w|$ and

$$|u| = |f_3| - |v_0v_1w_1'v_0v_1w_2| + |g_1v_0v_1t'|$$

and $|t'| \leq |z_0| \leq |z| < |bz| \leq |v_0v_1|$ and $|g_1| \leq |w_1'|$ and $w_2 \neq \varepsilon$. Assume, f_3 is bordered. Then f_3 has a shortest border h_3 such that $v_0v_1 \leq h_3$. We have $h_3 = v_0u_1''v_1$ by the arguments from the previous paragraph. Moreover,

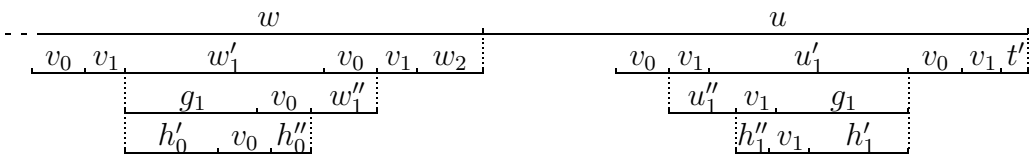
$$v_0v_1u_1' = h_3g_1 \qquad \text{and} \qquad v_0v_1w_1' \leq h_3 \ . \tag{8}$$



Observe, that (7) and (8) imply that the number of occurrences of v_1 and v_0, respectively, is the same in w_1' and u_1' since v_0 and v_1 do not overlap. Now, let

$$h_1 = v_1g_1v_0 = h_1''v_1h_1'v_0 = v_1h_0'v_0h_0''$$

where v_1 and v_0 occur only once in v_1h_1' and $h_0'v_0$, respectively.



Now, let

$$f_2' = v_0 h_0'' w_1'' v_1 w_2 u_0' v_0 v_1 u_j' \cdots v_0 v_1 u_1' v_0 v_1$$

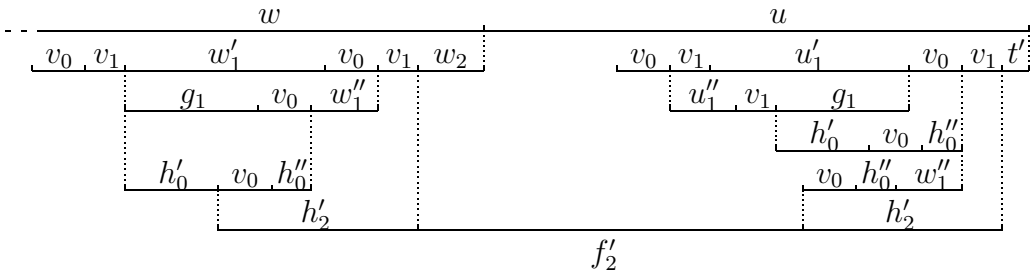
and

$$f_3' = v_0 v_1 w_1' v_0 v_1 w_2 u_0' v_0 v_1 u_j' \cdots v_0 v_1 u_2' v_0 u_1'' h_1'' v_1$$

with respective shortest borders h_2' and h_3' (which are both not empty, if $|u| \geq |w| - 1$; as in the case of f_2 and f_3) and $v_0 v_1 \preccurlyeq h_2'$ and $v_0 v_1 \leq h_3'$.

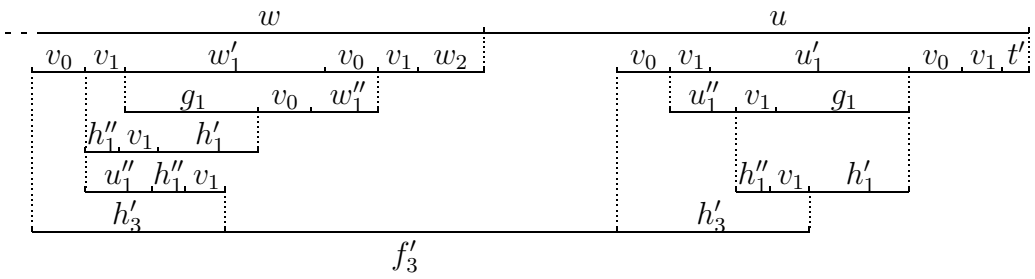
We have $h_2' \leq v_0 h_0'' w_1'' v_1$ since $v_0 v_1$ does not occur in w_2 and az does not occur in h_2' (and so h_2' does not stretch beyond w). We have $v_0 h_0'' w_1'' v_1 \leq h_2'$ since $v_0 v_1$ does not occur in w_1'. Hence, we have $h_2' = v_0 h_0'' w_1'' v_1$ and

$$w_1' v_0 v_1 = h_0' v_0 h_2'' w_1'' v_1 = h_0' h_2' \qquad \text{and} \qquad h_2' \preccurlyeq u_1' v_0 v_1 \ .$$



We have $h_3' = v_0 u_1'' h_1'' v_1$ by the arguments from the previous paragraph. Moreover,

$$v_0 v_1 u_1' = v_0 u_1'' h_1'' v_1 h_1' = h_3' h_1' \qquad \text{and} \qquad v_0 v_1 w_1' \leq h_3' \ .$$



It is now straightforward to see that

$$w_1'' = u_1'' = \varepsilon$$

for otherwise v_1 and v_0 occur more than once in $v_1 h_1'$ and $h_0' v_0$, respectively. From (6) follows now

$$w_1' = g_1 = u_1' \ .$$

Assume $1 < k \leq \min\{i, j\}$ and $w_\ell' = u_\ell'$, for all $1 \leq \ell < k$. Let us denote both w_ℓ' and u_ℓ' by v_ℓ', for all $1 \leq \ell < k$.

We show that $w'_k = u'_k$. Consider

$$f_4 = v_1 w'_k v_0 v_1 v'_{k-1} v_0 v_1 \cdots v'_1 v_0 v_1 w_2 u'_0 v_0 v_1 u'_j \cdots v_0 v_1 u'_k v_0 \ .$$

If f_4 is unbordered, then $|u| < |w| - 1$ since $|f_4| \leq |w|$ and

$$|u| = |f_4| - |v_1 w'_k v_0 v_1 v'_{k-1} v_0 v_1 \cdots v'_1 v_0 v_1 w_2| + |v_1 v'_{k-1} v_0 v_1 \cdots v'_1 v_0 v_1 t'|$$

and $|t'| \leq |z_0| \leq |z| < |bz| \leq |v_0 v_1|$ and $w_2 \neq \varepsilon$. Assume, f_4 is bordered. Then f_4 has a shortest border h_4 such that $|v_0 v_1| \leq |h_4|$. Let $h_4 = v_1 g_4 v_0$.

If $|v_1 w'_k v_0| < |h_4|$ then there exists an $\ell < k$ such that

$$h_4 = v_1 w'_k v_0 v_1 v'_{k-1} v_0 v_1 \cdots v'_{\ell+1} v_0 v_1 v''_\ell v_0$$

where $v''_\ell \leq v'_\ell$. That implies

$$u'_k = v''_\ell$$

since $v_0 v_1$ does neither occur in v''_ℓ nor in u'_k. Now, consider

$$f_5 = v_1 w'_k v_0 v_1 v'_{k-1} v_0 v_1 \cdots v'_1 v_0 v_1 w_2 u'_0 v_0 v_1 u'_j \cdots v_0 v_1 u'_k v_0 v_1 v'_{k-1} v_0 v_1 \cdots v''_\ell v_0 \ .$$

If f_5 is unbordered, then $|u| < |w| - 1$ since $|f_4| < |f_5|$, see above. Assume, f_5 is bordered. Then f_5 has a shortest border h_5 such that

$$|h_4| < |h_5|$$

for otherwise h_4 is not the shortest border of f_4, since either $h_4 \leq h_5$ or $h_5 \leq h_4$, and the latter implies that h_4 is bordered, and hence, not minimal. But now, we have a $\ell' < \ell$ such that

$$h_5 = v_1 w'_k v_0 v_1 v'_{k-1} v_0 v_1 \cdots v'_{\ell'+1} v_0 v_1 v''_{\ell'} v_0$$

where $v''_{\ell'} \leq v'_{\ell'}$. We have $|f_4| < |f_5| < |f_6|$ where

$$f_6 = v_1 w'_k v_0 v_1 v'_{k-1} v_0 v_1 \cdots v'_1 v_0 v_1 w_2 u'_0 v_0 v_1 u'_j \cdots v_0 v_1 u'_k v_0 v_1 v'_{k-1} v_0 v_1 \cdots v''_{\ell'} v_0 \ ,$$

which is either unbordered and $|u| < |w| - 1$ since $|f_4| < |f_5|$, or it is bordered with a shortest border h_6, and we have $|h_4| < |h_5| < |h_6|$ and a factor f_7, such that $|f_4| < |f_5| < |f_6| < |f_7|$, and so on, until eventually an unbordered factor is reached proving that $|u| < |w| - 1$.

Assume then that $h_4 \leq v_1 w'_k v_0$. We also have that $h_4 \preccurlyeq v_1 u'_k v_0$ since $v_0 v_1$ does not occur in w'_k. So, let $w'_k v_0 = g_4 v_0 w''_k$ and $v_1 u'_k = u''_k v_1 g_4$.

Consider,

$$f_8 = v_0 w''_k v_1 v'_{k-1} v_0 v_1 \cdots v'_1 v_0 v_1 w_2 u'_0 v_0 v_1 u'_j v_0 v_1 \cdots u'_k v_0 v_1 \ .$$

If f_8 is unbordered, then $|u| < |w| - 1$ since $|f_8| \leq |w|$ and

$$|u| = |f_8| - |v_0 w_k'' v_1 v_{k-1}' v_0 v_1 \cdots v_1' v_0 v_1 w_2| + |v_{k-1}' v_0 v_1 \cdots v_1' v_0 v_1 t'|$$

and $|t'| \leq |z_0| \leq |z| < |bz| \leq |v_0 v_1|$ and $w_2 \neq \varepsilon$. Assume, f_8 is bordered. Then f_8 has a shortest border h_8 such that $v_0 v_1 \preccurlyeq h_8$.

If $|h_8| > |v_0 w_k'' v_1|$ then the same argument as in the case $|v_1 w_k' v_0| < |h_4|$ above shows that $|u| < |w| - 1$. If $|h_8| < |v_0 w_k'' v_1|$ then $v_0 v_1$ occurs in w_k'; a contradiction. Hence, we have $h_8 = v_0 w_k'' v_1$ and

$$w_k' v_0 v_1 = g_1 h_8 \qquad \text{and} \qquad h_8 \preccurlyeq u_k' v_0 v_1 \ . \tag{9}$$

Consider,

$$f_9 = v_0 v_1 w_k' v_0 v_1 v_{k-1}' v_0 v_1 \cdots v_1' v_0 v_1 w_2 u_0' v_0 v_1 u_j' v_0 v_1 \cdots u_{k+1}' v_0 u_k'' v_1 \ .$$

If f_9 is unbordered, then $|u| < |w| - 1$ since $|f_9| \leq |w|$ and

$$|u| = |f_9| - |v_0 v_1 w_k' v_0 v_1 v_{k-1}' v_0 v_1 \cdots v_1' v_0 v_1 w_2| + |g_4 v_0 v_1 v_{k-1}' v_0 v_1 \cdots v_1' v_0 v_1 t'|$$

and $|t'| \leq |z_0| \leq |z| < |bz| \leq |v_0 v_1|$ and $|g_4| \leq |w_k'|$ and $w_2 \neq \varepsilon$. Assume, f_9 is bordered. Then f_9 has a shortest border h_9 such that $v_0 v_1 \leq h_9$. We have $h_9 = v_0 u_k'' v_1$ by the arguments from the previous paragraph. Moreover,

$$v_0 v_1 u_k' = h_9 g_1 \qquad \text{and} \qquad h_9 \leq v_0 v_1 w_k' \ . \tag{10}$$

Observe, that (9) and (10) imply that the number of occurrences of v_1 and v_0, respectively, is the same in w_k' and u_k' since v_0 and v_1 do not overlap. Now, let

$$h_4 = v_1 g_4 v_0 = h_1'' v_1 h_1' v_0 = v_1 h_0' v_0 h_0''$$

where v_1 and v_0 occur only once in $v_1 h_1'$ and $h_0' v_0$, respectively.

Now, let

$$f_8' = v_0 h_0'' w_k'' v_1 v_{k-1}' \cdots v_0 v_1 v_1' v_0 v_1 w_2 . u_0' v_0 v_1 u_j' \cdots v_0 v_1 u_k' v_0 v_1$$

and

$$f_9' = v_0 v_1 w_k' v_0 v_1 v_{k-1}' \cdots v_0 v_1 v_1' v_0 v_1 w_2 . u_0' v_0 v_1 u_j' \cdots v_0 v_1 u_{k+1}' v_0 u_1'' h_1'' v_1$$

with respective shortest borders h_8' and h_9' (which are both not empty, if $|u| \geq |w| - 1$; as in the case of f_8 and f_9). Analogously to the cases of f_8 and f_9, we have

$$w_k' v_0 v_1 = h_0' h_8' \qquad \text{and} \qquad v_0 v_1 u_k' = h_9' h_1' \ .$$

It is now straightforward to see that

$$h'_8 = h'_9 = v_0v_1$$

and

$$h_4 = v_0w'_kv_1 = v_0u'_kv_1$$

and hence, $w'_k = u'_k$. In this case, we denote both w'_k and u'_k by v'_k.

Now, we have

$$\begin{aligned}\bar{v} &= v_0v_1w'_\iota \cdots v_0v_1w'_2v_0v_1w'_1\\ &= v_0v_1u'_\iota \cdots v_0v_1u'_2v_0v_1u'_1\end{aligned}$$

where $\iota = \min\{i, j\}$.

If $i < j$ then

$$|w'_0| < |u'_0v_0v_1u'_j \cdots v_0v_1u'_{i+1}| \tag{11}$$

since $|w'_0| \leq |u'_0|$ by (5). Let

$$f_{11} = v_1w_2u'_0v_0v_1u'_j \cdots v_0v_1u'_{i+1}\bar{v}v_0 \ .$$

Then $|w| < |f_{11}|$ by (11), and hence, f_{11} is bordered. Let $h_{11} = v_1g_{11}v_0$ be the shortest border of f_{11}. Recall, that $w_2 \neq \varepsilon$ and either $az \preccurlyeq v_1w_2$ or $v_1w_2 \preccurlyeq az$. If $|v_1w_2| < |az|$ then v_1 necessarily occurs in z, and hence, it overlaps with v_0 (since $bz \leq v_0v_1$); a contradiction. So, we have $az \preccurlyeq v_1w_2$. Surely, $|h_{11}| < |v_1w_2|$ (and so $h_{11} \leq v_1w_2$) for otherwise az occurs in u which contradicts our assumption that z is of maximum length. Let $w_2 = g_{11}v_0w_5$. Note, that $|v_0w_5| \neq |az|$ since az and v_0 begin with different letters. We have $|az| < |v_0w_5|$ since otherwise v_0 occurs in z, and hence, overlaps with v_1 which is a contradiction. Consider,

$$f_{12} = v_0w_5u'_0v_0v_1u'_j \cdots v_0v_1u'_{i+1}\bar{v}v_0v_1 \ .$$

If f_{12} is unbordered, then $|u| < |w| - 1$ since $|f_{12}| \leq |w|$ and

$$|u| = |f_{12}| - |v_0w_5| + |t'|$$

and $|az| < |v_0w_5|$ and $|t'| \leq |z_0| \leq |z| < |bz| < |v_0w_5|$. Assume, f_{12} is bordered. Then f_{12} has a shortest border $h_{12} = g_{12}v_0v_1$ with $|az| < |h_{12}|$, for otherwise az occurs in u. Let $v_0w_5 = g_{12}v_0v_1w_6$. But, now

$$w = w'_0\bar{v}v_0v_1g_{12}v_0v_1w_6$$

where $v_0v_1w_6 \preccurlyeq w_2$, contradicting our assumption that v_0v_1 does not occur in w_2.

If $i > j$ then

$$w = w'_0 v_0 v_1 w'_i \cdots v_0 v_1 w'_{j+1} \bar{v} v_0 v_1 w_2 \qquad \text{and} \qquad u = u'_0 \bar{v} v_0 v_1 t'$$

and $|w| \geq |u| - |t'| + |v_0 v_1|$. We have $|u| < |w| - 1$ since $|t'| \leq |z_0| < |v_0 v_1| - 1$ by (3).

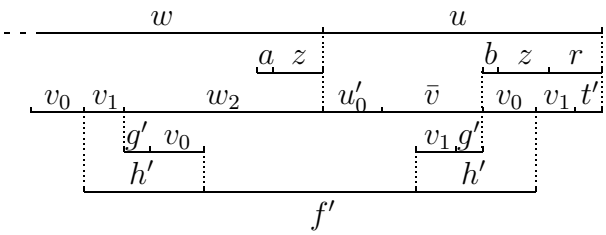
Assume $i = j$. Then

$$w = w'_0 \bar{v} v_0 v_1 w_2 \qquad \text{and} \qquad u = u'_0 \bar{v} v_0 v_1 t' \ .$$

Consider

$$f' = v_1 w_2 u'_0 \bar{v} v_0 \ .$$

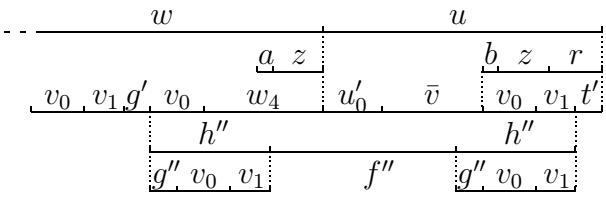
If f' is bordered, then it has a shortest border $h' = v_1 g' v_0$.



Recall, that $w_2 \neq \varepsilon$ and either $az \preccurlyeq v_1 w_2$ or $v_1 w_2 \preccurlyeq az$. If $|v_1 w_2| < |az|$ then v_1 occurs in z, and hence, overlaps with v_0 since $bz \leq v_0 v_1$; a contradiction. So, we have $az \preccurlyeq v_1 w_2$. Surely, $|h'| < |v_1 w_2|$ for otherwise az occurs in u which contradicts our assumption. Let $w_2 = g' v_0 w_4$. Note, that $|v_0 w_4| \neq |az|$ since az and v_0 begin with different letters. We have $|az| < |v_0 w_4|$ since otherwise v_0 occurs in z, and hence, overlaps with v_1 which is a contradiction. Consider now,

$$f'' = v_0 w_4 u'_0 \bar{v} v_0 v_1 \ .$$

If f'' is unbordered, then it easily follows that $|u| < |w| - 1$ since we have $|t'| < |az| < |v_0 w_4|$.



If f'' is bordered, then it has a shortest border $h'' = g'' v_0 v_1$ with $|az| < |h''|$, for otherwise az occurs in u. Let $v_0 w_4 = g'' v_0 v_1 w_5$. But, now

$$w = w'_0 \bar{v} v_0 v_1 g' g'' v_0 v_1 w_5$$

which contradicts our assumption that $w = w'_0\bar{v}v_0v_1w_2$ and v_0v_1 does not occur in w_2.

If f' is unbordered, then $|f'| \leq |w|$, and hence, $|w'_0| \geq |u'_0|$. But, we also have $|w'_0| \leq |u'_0|$; see (5). That implies $|w'_0| = |u'_0|$. Moreover, the factors w_0 and bzv' have both nonoverlaping occurrences in $u'_0v_0v_1$ by (5). Therefore, $w'_0 = u'_0$. Now,

$$w = xaw_7 \qquad \text{and} \qquad u = xbt''$$

where $w'_0\bar{v}v_0v_1 \leq x$ and $a, b \in A$ and $a \neq b$ and $w_7 \preccurlyeq w_2$ and $t'' \preccurlyeq t'$. We have that xb occurs in w by Theorem 11. Since xb is not a prefix of w and v_0v_1 does not overlap with itself, we have $|xb| + |v_0v_1| \leq |w|$. From $|t'| \leq |z_0| < |v_0v_1| - 1$ we get $|u| < |w| - 1$ and the claim follows. □

Note, that the bound $|u| < |w| - 1$ on the length of a nontrivial Duval extension wu of w is tight, as the example given in the introduction shows. Theorem 2 also implies a new bound on the length of any word w such that $\partial(w) = \mu(w)$ must hold.

Corollary 3. *If* $|w| \geq 3\mu(w) - 2$ *then* $\partial(w) = \mu(w)$.

5 Conclusions

In this paper we have given a confirmative answer to a long standing conjecture [10] by proving that a Duval extension wu of w longer than $2|w| - 2$ is trivial. This bound is thight and also gives a new bound on the relation between the length of an arbitrary word w and its longest unbordered factors $\mu(w)$, namely that $|w| \geq 3\mu(w) - 2$ implies $\partial(w) = \mu(w)$ as conjectured (more weakly) in [1]. We believe that the precise bound can be achieved with methods similar to those presented in this paper.

References

[1] R. Assous and M. Pouzet. Une caractérisation des mots périodiques. *Discrete Math.*, 25(1):1–5, 1979.

[2] J. Berstel and D. Perrin. *Theory of codes*, volume 117 of *Pure and Applied Mathematics*. Academic Press Inc., Orlando, FL, 1985.

[3] R. S. Boyer and J. S. Moore. A fast string searching algorithm. *Commun. ACM*, 20(10):762–772, October 1977.

[4] D. Breslauer, T. Jiang, and Z. Jiang. Rotations of periodic strings and short superstrings. *J. Algorithms*, 24(2), 1997.

[5] P. Bylanski and D. G. W. Ingram. *Digital transmission systems*. IEE, 1980.

[6] Y. Césari and M. Vincent. Une caractérisation des mots périodiques. *C. R. Acad. Sci. Paris Sér. A*, 286:1175–1177, 1978.

[7] M. Crochemore, F. Mignosi, A. Restivo, and S. Salemi. Text compression using antidictionaries. In *26th Internationale Colloquium on Automata, Languages and Programming (ICALP), Prague*, volume 1644 of *Lecture Notes in Comput. Sci.*, pages 261–270. Springer, Berlin, 1999.

[8] M. Crochemore and D. Perrin. Two-way string-matching. *J. ACM*, 38(3):651–675, 1991.

[9] J.-P. Duval. Périodes et répétitions des mots de monoïde libre. *Theoret. Comput. Sci.*, 9(1):17–26, 1979.

[10] J.-P. Duval. Relationship between the period of a finite word and the length of its unbordered segments. *Discrete Math.*, 40(1):31–44, 1982.

[11] A. Ehrenfeucht and D. M. Silberger. Periodicity and unbordered segments of words. *Discrete Math.*, 26(2):101–109, 1979.

[12] T. Harju and D. Nowotka. Density of critical factorizations. *Theor. Inform. Appl.*, 36(3):315–327, 2002.

[13] T. Harju and D. Nowotka. Duval's conjecture and Lyndon words. technical report 479, Turku Centre of Computer Science (TUCS), Turku, Finland, October 2002. submitted.

[14] T. Harju and D. Nowotka. Minimal Duval extensions. technical report 520, Turku Centre of Computer Science (TUCS), Turku, Finland, April 2003. submitted.

[15] D. E. Knuth, J. H. Morris, and V. R. Pratt. Fast pattern matching in strings. *SIAM J. Comput.*, 6(2):323–350, 1977.

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

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

[18] R. C. Lyndon. On Burnside's problem. *Trans. Amer. Math. Soc.*, 77:202–215, 1954.

[19] D. Margaritis and S. Skiena. Reconstructing strings from substrings in rounds. In *36th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 613–620, Milwaukee, WI, 1995. IEEE Computer Society.

[20] F. Mignosi and L. Q. Zamboni. A note on a conjecture of Duval and Sturmian words. *Theor. Inform. Appl.*, 36(1):1–3, 2002.

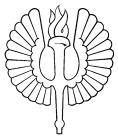
[21] M. Morse and G. A. Hedlund. Symbolic dynamics II: Sturmian trajectories. *Amer. J. Math.*, 61:1–42, 1940.

[22] M.-P. Schützenberger. A property of finitely generated submonoids of free monoids. In *Algebraic theory of semigroups (Proc. Sixth Algebraic Conf., Szeged, 1976)*, volume 20 of *Colloq. Math. Soc. János Bolyai*, pages 545–576. North-Holland, Amsterdam, 1979.

[23] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Information Theory*, 23(3):337–343, 1977.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.fi



University of Turku
- **Department of Information Technology**
- **Department of Mathematics**

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

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**