The Ehrenfeucht–Silberger Problem[☆]

Štěpán Holub[a,1], Dirk Nowotka[*,b]

[a] *Department of Algebra, Charles University of Prague, Sokolovska 83, 186 75 Praha 8, Czech Republic.*
[b] *Institute for Formal Methods in Computer Science, Universität Stuttgart, Universitätsstr. 38, 70569 Stuttgart, Germany.*

Abstract

We consider repetitions in words and solve a longstanding open problem about the relation between the period of a word and the length of its longest unbordered factor (where factor means uninterrupted subword). A word u is called bordered if there exists a proper prefix that is also a suffix of u, otherwise it is called unbordered. In 1979 Ehrenfeucht and Silberger raised the following problem: What is the maximum length of a word w, w.r.t. the length τ of its longest unbordered factor, such that τ is shorter than the period π of w. We show that, if w is of length $\frac{7}{3}\tau$ or more, then $\tau = \pi$ which gives the optimal asymtotic bound.

Key words: combinatorics on words, Ehrenfeucht–Silberger problem, periodicity, unbordered words

1. Introduction

When repetitions in words are considered then two notions are central: the *period*, which gives the least amount by which a word has to be shifted in order to overlap with itself, and the shortest *border*, which denotes the least (nontrivial) overlap of a word with itself. Both notions are related in several ways, for example,

[☆] A preliminary version of this paper appeared in the proceedings of the ICALP 2009.

[*] Corresponding author

Email addresses: holub@karlin.mff.cuni.cz (Štěpán Holub), nowotka@fmi.uni-stuttgart.de (Dirk Nowotka)

[1] The work on this article has been supported by the research project MSM 0021620839.

the period of an unbordered word is its length, and the length of the shortest border of a bordered word is not larger than its period. Moreover, a shortest border itself is always unbordered. Periodicity also restricts occurrences of long unbordered factors (that is, uninterrupted substrings). Deeper dependencies between the period of a word and its unbordered factors have been investigated and exploited in applications for decades; see also references to related work below.

Let us recall the problem by Ehrenfeucht and Silberger. Let w be a (finite) word of length $|w|$, let $\tau(w)$ denote the maximum length of unbordered factors of w, and let $\pi(w)$ denote the period of w. Certainly, $\tau(w) \leq \pi(w)$ since a period of w is also a period of its factors. Moreover, it is well-known that $\tau(w) = \pi(w)$ when $|w| \geq 2\pi(w)$. So, the interesting cases are those where $|w| < 2\pi(w)$. Actually, the interesting cases are also the most common ones since by far most words have a period that is longer than one half of their length. When such words are considered, a bound on $|w|$, enforcing $\tau(w) = \pi(w)$, that depends on $\tau(w)$ becomes more interesting than the one depending on $\pi(w)$.

The problem by Ehrenfeucht and Silberger asks about a bound on $|w|$ depending on $\tau(w)$ such that $\tau(w) = \pi(w)$ is enforced. In this paper we establish the following fact for all finite words w:

$$\text{If} \quad |w| \geq \frac{7}{3}\tau(w) \quad \text{then} \quad \tau(w) = \pi(w)\,.$$

This bound on the length of w is asymtotically tight; see the following example by Assous and Pouzet.

Previous Work

Ehrenfeucht and Silberger raised the problem described above in [1]. They conjectured that $|w| \geq 2\,\tau(w)$ implies $\tau(w) = \pi(w)$. That conjecture was falsified shortly thereafter by Assous and Pouzet [2] by the following example:

$$w = a^n b a^{n+1} b a^n b a^{n+2} b a^n b a^{n+1} b a^n \tag{1}$$

where $n \geq 0$ and $\tau(w) = 3n+6$ (note that $ba^{n+1}ba^nba^{n+2}$ and $a^{n+2}ba^nba^{n+1}b$ are the two longest unbordered factors of w) and $\pi(w) = 4n+7$ and $|w| = 7n+10$, that is, $\tau(w) < \pi(w)$ and $|w| = 7/3\,\tau(w) - 4 > 2\tau(w)$. Assous and Pouzet in turn conjectured that $3\tau(w)$ is the bound on the length of w for establishing $\tau(w) = \pi(w)$. Duval [3] did the next step towards answering the conjecture. He established that $|w| \geq 4\,\tau(w) - 6$ implies $\tau(w) = \pi(w)$ and conjectured that, if w possesses an unbordered prefix of length $\tau(w)$, then $|w| \geq 2\,\tau(w)$ implies $\tau(w) = \pi(w)$. Despite some partial results [4, 5, 6] towards a solution, Duval's conjecture was only solved in 2004 [7, 8] with a new proof given in [9]. It turned out that the optimal bound is $2\,\tau(w) - 1$. Note that a positive answer to (the extended version of) Duval's conjecture lowered the bound for Ehrenfeucht and Silberger's problem to $3\,\tau(w) - 2$, in accordance with the conjecture by Assous and Pouzet [2].

However, there remained a gap of $\tau(w)/3$ between that bound and the largest known example which is given above. The bound of $\frac{7}{3}\tau(w)$ has been conjectured in [7, 8]. This conjecture is proved here, and the problem by Ehrenfeucht and Silberger is finally solved.

Other Related Work

The result related most closely to the problem by Ehrenfeucht and Silberger is the so called critical factorization theorem (CFT).

The CFT states the following: Let $w = uv$ be a factorization of a word w into u and v. The local period of w at the point $|u|$ is the length q of the shortest square centered at $|u|$. More formally, let x be the shortest word such that x is a prefix of vy and a suffix of zu for some y and z; then $q = |x|$. It is straightforward to see that q is not larger than the period of w. The factorization uv is called critical if q equals the period of w. The CFT states that a critical factorization exists for every nonempty word w, and moreover, a critical factorization uv can always be found such that $|u|$ is shorter than the period of w. The CFT was conjectured first by Schützenberger [10], proved by Césari and Vincent [11], and brought into its current form by Duval [12]. Crochemore and Perrin [13] found a new elegant proof of the CFT using lexicographic orders, and realized a direct application of the theorem in a new string-matching algorithm.

How does the CFT relate to the problem by Ehrenfeucht and Silberger? Observe that the shortest square x^2 centered at some point in w is always such that x is unbordered. If x results from a critical factorization and occurs in w (recall that x can, by definition, outreach the limits of w), then w contains an unbordered factor of the length of its period. Therefore, it follows from the CFT that $|w| > 2\pi(w) - 2$ implies $\tau(w) = \pi(w)$. This bound is asymptotically optimal. In this paper, we establish the asymptotically optimal bound on $|w|$ enforcing the equality $\tau(w) = \pi(w)$ in terms of $\tau(w)$ instead of $\pi(w)$. This rounds off the long lasting research effort on the mutual relationship between the two basic properties of a word w, that is, $\tau(w)$ and $\pi(w)$.

2. Notation and Basic Facts

Let us fix a finite set A of letters, called alphabet, for the rest of this paper. Let A^* denote the monoid of all finite words over A including the *empty word* denoted by ε. Let $w = uv \in A^*$. Then $u^{-1}w = v$ and $wv^{-1} = u$. In general, we denote variables over A by a, b, c, and d and variables over A^* are usually denoted by f, g, h, r through z, and by Greek letters, including their subscripted and primed versions. The letters i through q are to range over the set of nonnegative integers.

Let $w = a_1a_2 \cdots a_n$. We denote the length n of w by $|w|$, in particular $|\varepsilon| = 0$. Let $1 \leq i \leq j \leq n$. Then $u = a_ia_{i+1} \cdots a_j$ is called a *factor* of w. A factor u is called *proper* when $u \neq w$, that is, $i \neq 0$ or $j \neq n$. Let $0 \leq i \leq n$. Then $u = a_1a_2 \cdots a_i$ is called a *prefix* of w, denoted by $u \leq_{\mathrm{p}} w$, and $v = a_{i+1}a_{i+2} \cdots a_n$ is called a *suffix* of w, denoted by $v \leq_{\mathrm{s}} w$. A prefix or suffix is called *proper* when $0 < i < n$. The *longest common prefix* w of two words u and v is denoted by $u \wedge_{\mathrm{p}} v$ and is defined by $w = u$, if $u \leq_{\mathrm{p}} v$, or $w = v$, if $v \leq_{\mathrm{p}} u$, or $wa \leq_{\mathrm{p}} u$ and $wb \leq_{\mathrm{p}} v$ for some different letters a and b. The *longest common suffix* of u and v, denoted $u \wedge_{\mathrm{s}} v$, is defined similarly, as one would expect. Two words u and

v, with $|u| \leq |v|$, *overlap* each other, if there is a word w, with $|v| < |w| < |uv|$, such that $u \leq_{\mathrm{p}} w$ and $v \leq_{\mathrm{s}} w$ or $v \leq_{\mathrm{p}} w$ and $u \leq_{\mathrm{s}} w$. An integer $1 \leq p \leq n$ is a *period* of w if $a_i = a_{i+p}$ for all $1 \leq i \leq n - p$. The smallest period of w is called *the period* of w, denoted by $\pi(w)$. A nonempty word u is called a *border* of a word w, if $w = uy = zu$ for some nonempty words y and z. We call w *bordered*, if it has a border, otherwise w is called *unbordered*. Let $\tau(w)$ denote the maximum length of unbordered factors of w, and $\tau_2(w)$ denote the maximum length of unbordered factors occurring at least twice in w. We have that

$$\tau(w) \leq \pi(w)\,. \tag{2}$$

Indeed, let $u = b_1 b_2 \cdots b_{\tau(w)}$ be an unbordered factor of w. If $\tau(w) > \pi(w)$ then $b_i = b_{i+\pi(w)}$ for all $1 \leq i \leq \tau(w) - \pi(w)$ and $b_1 b_2 \cdots b_{\tau(w)-\pi(w)}$ is a border of u; a contradiction.

Let $\lhd$ be a total order on A. Then $\lhd$ extends to a *lexicographic order*, also denoted by $\lhd$, on A^* with $u \lhd v$ if either $u \leq_{\mathrm{p}} v$ or $xa \leq_{\mathrm{p}} u$ and $xb \leq_{\mathrm{p}} v$ and $a \lhd b$. Let $\lhd^a$ denote a lexicographic order where the maximum letter a is fixed in the respective order on A. A $\lhd$-*maximum suffix* α of a word w is defined as a suffix of w such that $v \lhd \alpha$ for all $v \leq_{\mathrm{s}} w$.

The following remarks state some facts about maximum suffixes which are folklore. They are included in this paper to make it self-contained.

Remark 2.1. Let w be a bordered word. The shortest border u of w is unbordered, and $w = uzu$. The longest border of w has length equal to $|w| - \pi(w)$.

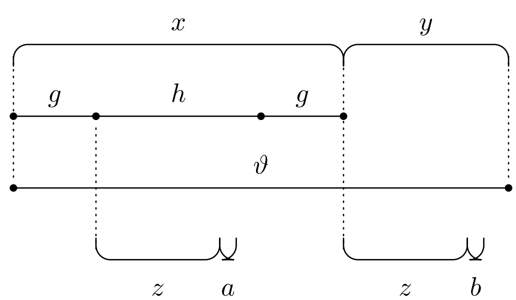
Indeed, if u is a border of w, then each border of u is also a border of w. Therefore u is unbordered, and it does not overlap with itself. If v is a border of w then $|w| - |v|$ is a period of w. Conversely, the prefix of w of length $|w| - \pi(w)$ is a border of w.

Remark 2.2. Any maximum suffix of a word w occurs only once in w and is longer than $|w| - \pi(w)$.

Indeed, let α be the $\lhd$-maximum suffix of w for some order $\lhd$. Then $w = x\alpha y$ and $\alpha \lhd \alpha y$ implies $y = \varepsilon$ by the maximality of α. If $w = uv\alpha$ with $|v| = \pi(w)$, then $u\alpha \leq_{\mathrm{p}} w$ gives a contradiction again.

Remark 2.3. Let ϑ be its own maximum suffix w.r.t. some order $\lhd$, and let x be a prefix of ϑ of length $\pi(\vartheta)$. Then x is unbordered.

Indeed, suppose on the contrary that x is bordered, that is, $x = ghg$ for some nonempty g. Let $\vartheta = xy$. We have $gy \lhd \vartheta = ghgy$, by assumption, which implies $y \lhd hgy$. Note that gy is not a prefix of ϑ otherwise $|gh| < |x|$ is a period of ϑ contradicting the choice of x. Hence, $zb \leq_{\mathrm{p}} y$ and $za \leq_{\mathrm{p}} hgy$ for some different letters a and b with $b \lhd a$. But, $y \leq_{\mathrm{p}} \vartheta$, since $|x| = \pi(w)$, implies $zb \leq_{\mathrm{p}} \vartheta$ which contradicts the maximality of ϑ (since $zb \leq_{\mathrm{p}} \vartheta \lhd za \leq_{\mathrm{p}} hgy$). These arguments are illustrated by the following figure.



Let an integer q with $0 \leq q < |w|$ be called a *point* in w. A nonempty word x is called a *repetition word at point* q if $w = uv$ with $|u| = q$ and there exist words y and z such that $x \leq_{\mathrm{s}} yu$ and $x \leq_{\mathrm{p}} vz$. Let $\pi(w, q)$ denote the length of the shortest repetition word at point q in w. We call $\pi(w, q)$ the *local period at point* q in w. Note that the repetition word of length $\pi(w, q)$ at point q is necessarily unbordered and $\pi(w, q) \leq \pi(w)$. A factorization $w = uv$, with $u, v \neq \varepsilon$ and $|u| = q$, is called *critical*, if $\pi(w, q) = \pi(w)$, and if this holds, then q is called a *critical point*. Let $\lhd$ be an order on A and $\blacktriangleleft$ be its inverse. Then the shorter of the $\lhd$-maximum suffix and the $\blacktriangleleft$-maximum suffix of some word w is called a *critical suffix* of w. This terminology is justified by the following version of the so called critical factorization theorem (CFT) [13] which relates maximum suffixes and critical points.

Theorem 2.4 (CFT). *Let w be a nonempty word and γ be a critical suffix of w. Then $|w| - |\gamma|$ is a critical point.*

Remark 2.5. Let rs be an unbordered word where $|r|$ is a critical point. Then s and r do not overlap and sr is unbordered with $|s|$ as a critical point.

3. Special Factorizations

Let us highlight the following definitions. They are not standard and will be central to the proof of Theorem 4.1. Let the words α and w be given.

Definition 3.1. The longest prefix of α strictly shorter than α that is also a suffix of w will be called the *α-suffix of w*.

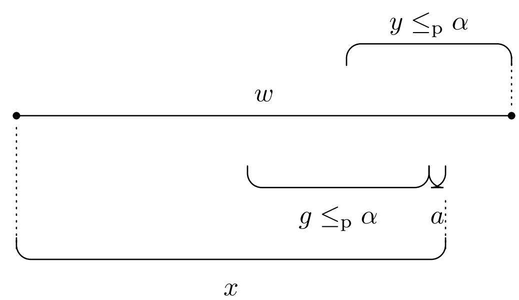
Definition 3.2. The number $|wy^{-1}|$, where y is the α-suffix of w, is called the *α-period of w*, denoted by $\pi_\alpha(w)$.

In particular, $|w| - |\alpha| < \pi_\alpha(w) \leq |w|$.

Definition 3.3. The shortest prefix x of w satisfying $\pi_\alpha(x) = \pi_\alpha(w)$ is called the *α-critical prefix of w*.

Remark 3.4. Note that the α-suffix of w can be empty, but it cannot be equal to α. For example, the *abb*-suffix of *aabb* is empty. Therefore, the *abb*-critical prefix of *aabb* is *aabb* itself. In general, if α is unbordered and it is a suffix of w, then the α-suffix of w is empty.

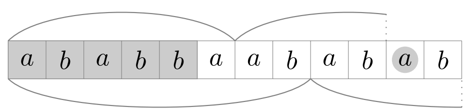
Let x be the α-critical prefix and y the α-suffix of w. Note that $\pi_\alpha(w) \leq |x| \leq |w|$ and, in particular, $\pi_\alpha(w) = |x| = |w|$ if $y = \varepsilon$. Consider the following illustration of the definitions with $ga \not\leq_{\mathrm{p}} \alpha$.



Remark 3.5. Note that $za = x$, where a is a letter, is the α-critical prefix of w if and only if za is the longest prefix of w satisfying $\pi_\alpha(z) < \pi_\alpha(za)$.

Example 3.6. Consider $w = ababbaababab$ of length 12 and $\alpha = ababb$. The α-suffix of w is $abab$, whence $\pi_\alpha(w) = 8$. The α-critical prefix of w is $ababbaababa$ of length 11, since

$$\pi_g(ababbaababa) = 8\,, \quad \text{and} \quad \pi_g(ababbaabab) = 6\,.$$



4. Solution of the Ehrenfeucht–Silberger Problem

This entire section is devoted to the proof of the main result of this paper: the solution of the Ehrenfeucht–Silberger problem by Theorem 4.1.

Theorem 4.1. *Let $w \in A^*$. If $|w| \geq \frac{7}{3}\tau(w)$ then $\tau(w) = \pi(w)$.*

We identify two particular unbordered factors of w and show that the assumption of the theorem, namely that these factors are strictly smaller than $\frac{3}{7}|w|$, leads to $\tau(w) = \pi(w)$.

Note that the claim holds trivially if every letter in w occurs only once because in that case $\tau(w) = \pi(w) = |w|$. We now define a factorization of w which is of central importance to our approach. Let

$$w = v'uzuv$$

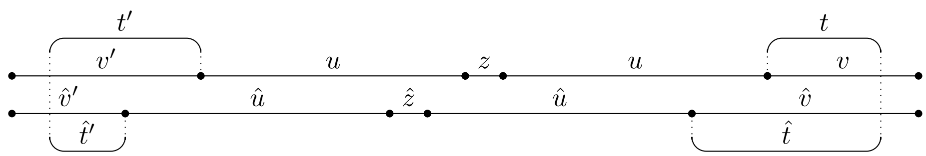
such that $|u| = \tau_2(w)$ and z is of maximum length (recall that $\tau_2(w)$ denotes the maximum length of unbordered factors occurring at least *twice* in w). Moreover, let us fix

$$t = v \wedge_{\mathrm{p}} zu \quad \text{and} \quad t' = v' \wedge_{\mathrm{s}} uz$$

for the rest of this proof.

It is clear that such a factorization exists whenever a letter occurs more than once in w. However, it is not necessarily unique. Suppose, for instance, that $t'u$ contains an unbordered factor $\hat{u}$, distinct from u but of the same length. Then we have a factorization $\hat{v}'\hat{u}\hat{z}\hat{u}\hat{v}$ of w, which also satisfies the requirements. Note, moreover, that if we define $\hat{t}$ and $\hat{t}'$ analogously to t and t', then we have

$$t^{-1}v = \hat{t}^{-1}\hat{v} \quad \text{and} \quad v't'^{-1} = \hat{v}'(\hat{t}')^{-1}. \tag{3}$$

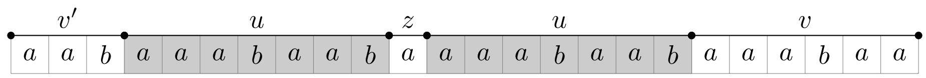


In one case it will be important to require that $t'u$ does not contain such an unbordered factor $\hat{u}$. That is, we shall single out the leftmost possible factorization (within bounds given by the factor $t'uzut$). We shall refer to this condition by saying that t' *is as short as possible.* If this additional assumption is not stated explicitly, then we consider an arbitrary factorization maximazing $|u|$ and $|z|$. The assumption is helpful in view of the following claim.

Claim 4.2. *Let t' be as short as possible, and let ϑ be a maximum suffix of $t'u$ w.r.t. some order $\lhd$. Then $|\vartheta| \leq |u|$.*

PROOF. Suppose that there is a maximum suffix ϑ of $t'u$ strictly longer than u. The prefix $\hat{u}$ of ϑ of length $\pi(\vartheta)$ is unbordered by Remark 2.3. It is of length at least $|u|$, since otherwise u is bordered. From $|u| = \tau_2(w)$ the equality $|\hat{u}| = |u|$ follows since $\hat{u}$ occurs at least twice in w; a contradiction with the minimality of t'.

The example of long words where the period exceeds the length of the longest unbordered factors by Assous and Pouzet (see page 3) turns out to highlight the most interesting cases of this proof. We therefore use its instance with $n = 2$ as a running example throughout this section. With this example a factorization of the above kind is illustrated by the following figure.



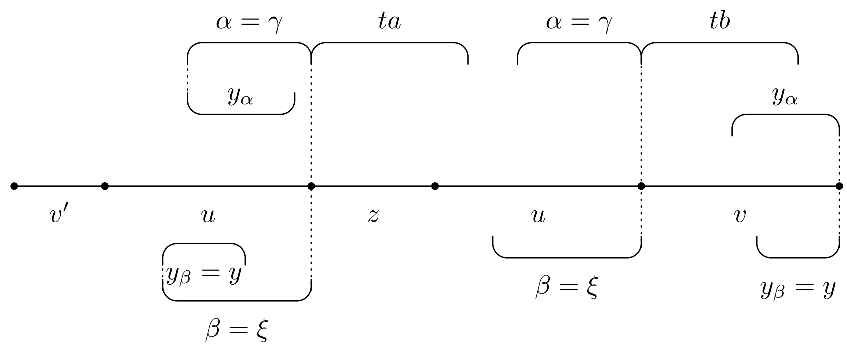
We start the proof by the following claim, which investigates a special situation, easy to exclude.

Claim 4.3. *Let ϑ be the maximum suffix of u w.r.t. some order $\lhd$. If $v_0\vartheta$ is a prefix of ϑv for some nonempty word v_0, then $uzu\vartheta^{-1}v_0\vartheta$ is unborderd.*

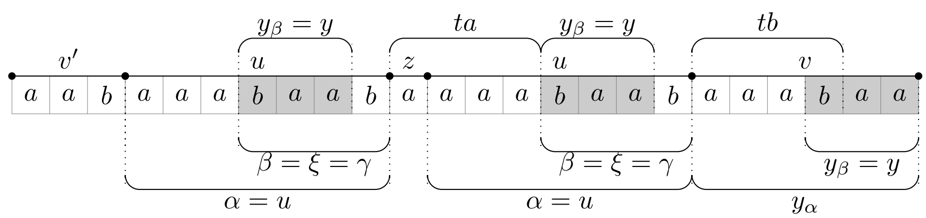
Proof. Suppose on the contrary that $uzu\vartheta^{-1}v_0\vartheta$ has a shortest border h. Note that h is, like every shortest border of a factor in w, not longer than $|u| = \tau_2(w)$. In fact $|h| < |u|$ since $|h| = |u|$ contradicts the maximality of $|z|$. If $|\vartheta| < |h| < |u|$ then ϑ occurs more than once in u contradicting Remark 2.2, which states that a maximum suffix occurs only once in a word. And finally, if $|h| \leq |\vartheta|$ then u is bordered by h since then $h \leq_{\mathrm{s}} \vartheta \leq_{\mathrm{s}} u$; a contradiction which concludes the proof.

As the reader already noted, our main tool will be considering maximum suffixes w.r.t. certain lexicographic orders. Let us therefore fix an order $\lhd$. Let α denote the $\lhd$-maximum suffix of u and β the $\blacktriangleleft$-maximum suffix of u, where $\blacktriangleleft$ is the inverse order of $\lhd$. Let y_α and y_β denote the α- and β-suffix of uv. Moreover, let y be the shorter of y_α and y_β and let ξ be either α or β so that $y = y_\xi$. Let γ denote the shorter of α and β. Note that $|y| < |\gamma|$ in any case.

The following figure shall illustrate the considered setting by an example where $v \neq t$ and $|\alpha| < |\beta|$ and $|y_\alpha| > |y_\beta|$, that is, we have $y = y_\beta$ and $\xi = \beta$ and $\gamma = \alpha$.



The same notation for our running example is depicted next.



It turns out that the proof splits into two main situations according to whether or not $|v| > |ty|$. Each of the cases yields a long unbordered factor of w.

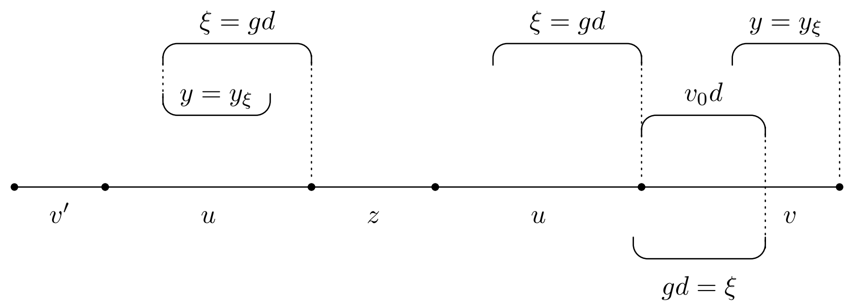
4.1. The First Factor

In this subsection we shall suppose $|v| > |ty|$ and consider the ξ-critical prefix of w. Note that the following claim holds independently of whether or not $v \neq t$.

Claim 4.4. *If* $|v| > |ty|$, *then* $\tau(w) \geq |\gamma zuvy^{-1}|$.

Proof. Supose $|v| > |ty|$. The inequality implies that the ξ-critical prefix of w can be written as $v'uzuv_0d$, where d is a letter. Let g denote the ξ-suffix of $v'uzuv_0$.

Assume first that $gd = \xi$ as illustrated by the next figure.

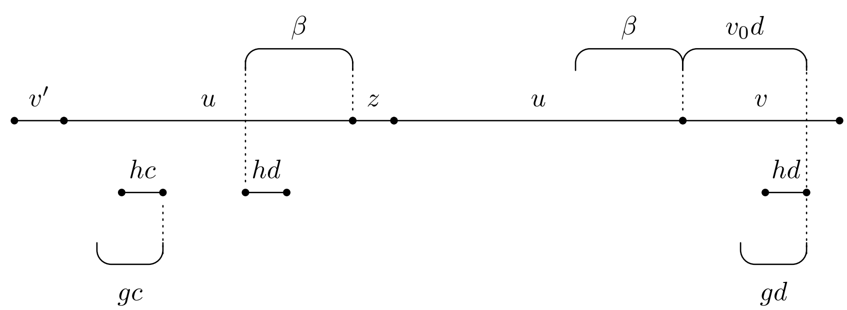


Then the word $uzuv_0d$ is unbordered, by Claim 4.3. Recall that $|\gamma| \leq |\xi| \leq |u|$ and that $|v_0d| \geq |vy^{-1}|$, since $v'uzuv_0d$ is the ξ-critical prefix of w. Therefore we have $\tau(w) \geq |uzuv_0d| \geq |\gamma zuvy^{-1}|$ as claimed.

Suppose next gc is a prefix of ξ with $c \neq d$. (Note that if $gd \neq \xi$, then $c \neq d$ is implied by the definition of the ξ-critical prefix.) We distinguish two cases on the order of c and d in $\lhd$.

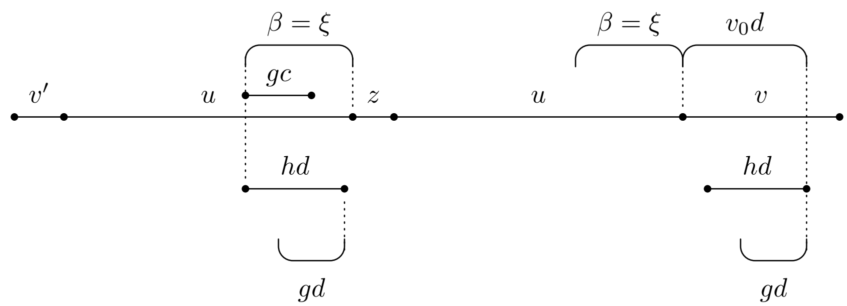
Suppose $c \lhd d$ and consider βzuv_0d. Recall that $|\beta| > |y|$ and $|v| \leq |v_0d| + |y|$. Hence, either βzuv_0d is unbordered and we get $\tau(w) \geq |\beta zuv_0d| \geq |\gamma zuvy^{-1}|$ and we are done, or βzuv_0d has a shortest border hd.

Suppose $|h| \leq |g|$ and $|h| < |\beta|$ as illustrated by the next figure.



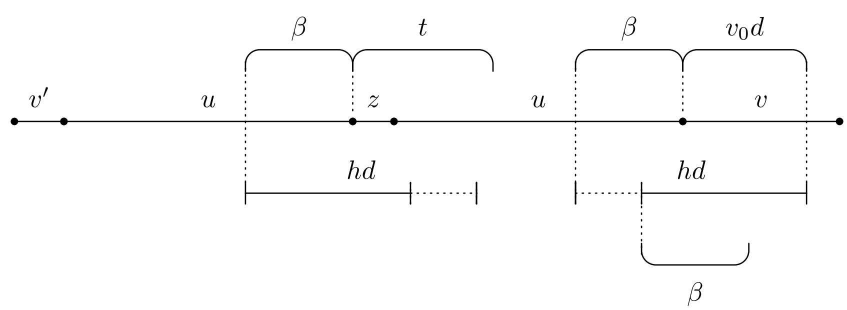
Then hd is a prefix of β and the occurrence of $hc \leq_{\mathrm{s}} gc$ in ξ, and hence also in u, contradicts the maximality of β since $hd \blacktriangleleft hc$.

Suppose $|g| < |h| < |\beta|$ as illustrated by the next figure.



Then gd occurs in u and $\xi = \beta$. Indeed, $gc \leq_{\mathrm{p}} \xi$ gives a contradiction if $\xi = \alpha$ since $gc \lhd gd$. But now, h contradicts the assumption that g is the ξ-suffix of $v'uzuv_0$.

It remains that $|\beta| \leq |h|$ which implies $\beta \leq_{\mathrm{p}} h$ as illustrated next.



The choice of u implies $|h| < |u|$. Hence, either $h = \beta v_0$ or the word $uzuv_0h^{-1}\beta$ is unbordered, by Claim 4.3. If $uzuv_0h^{-1}\beta$ is unbordered, then $|u| > |hd|$ and $|v| \leq |v_0d| + |y|$ imply $\tau(w) \geq |uzuv_0h^{-1}\beta| > |\beta zuv_0d| \geq |\gamma zuvy^{-1}|$. If $uzuv_0h^{-1}\beta$ is bordered, then $h = \beta v_0$, which implies $v_0d \leq_{\mathrm{p}} t$ (recall that $t = v \wedge_{\mathrm{p}} zu$), and $|v| \leq |ty|$, since $|v| \leq |v_0d| + |y|$; a contradiction. This completes the case $c \lhd d$.

The case $d \lhd c$ is similar considering αzuv_0d and the claim is thereby proved.

Remark 4.5. Note that we have arguments for v' symmetric to those for v. That is, if we define α', β', y', ξ' and γ' for v' analogously, then Claim 4.4 implies the following: If $|v'| > |t'y'|$, then $\tau(w) \geq |{y'}^{-1}v'uz\gamma'|$.

4.2. The Second Factor

In this section, we investigate the possibility $|v| \le |ty|$. We shall also suppose that v is not a prefix of zu, that is, $t \neq v$. In the rest of the paper, whenever $t \neq v$, the first letter of $t^{-1}v$ will be denoted by b and the first letter of $t^{-1}zu$ by a. In other words, the word ta is a prefix of zuv and tb a prefix of v, with $a \neq b$.

Let δ denote the word such that δa is the $\lhd^a$-maximum suffix of $t'uta$ for some fixed order $\lhd^a$ such that a is the maximum in A. The word δ plays an important role in this section, similar to the role of ξ in the previous section. We first point out that every factor of $t'uv$ is strictly less than δa w.r.t. $\lhd^a$ if $|v| \le |ty|$. In particular, δa does not occur in $t'uv$ in such a case.

Claim 4.6. *Let f be a factor of $t'uv$. If $|v| \le |ty|$, then $f \lhd^a \delta a$ and $f \neq \delta a$.*

Proof. If f occurs in $t'ut$ or y, then the claim follows from the maximality of δa.

Assuming $|v| \le |ty|$, it remains that there is a prefix $f'b$ of f such that $f' \le_{\mathrm{s}} t'ut$. Then $f'a \le_{\mathrm{s}} t'uta$, and the maximality of δa implies $f'a \lhd^a \delta a$. The claim now follows from $f'b \le_{\mathrm{p}} f$ and $f'b \lhd^a f'a$.

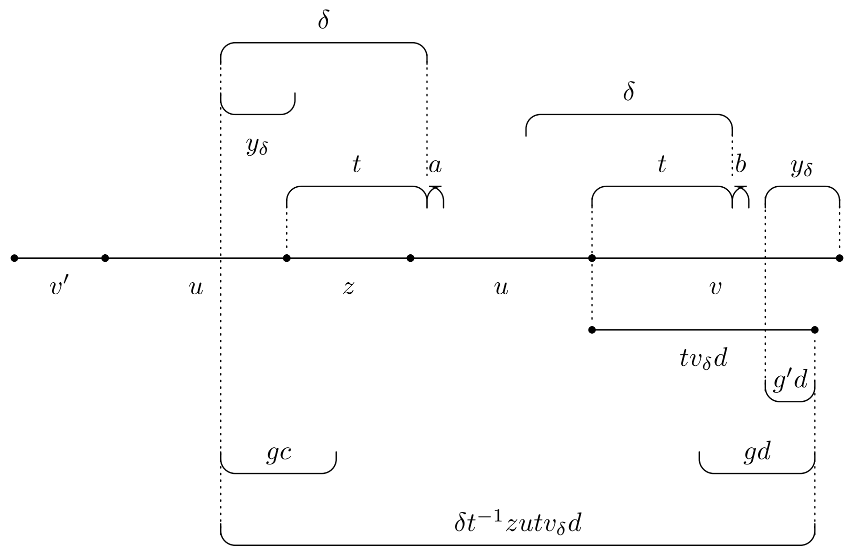
The following claim introduces a further long unbordered factor of w, namely $\delta t^{-1}zuvy_\delta^{-1}$, where y_δ is the δa-suffix of w.

Claim 4.7. *The word $\delta t^{-1}zuvy_\delta^{-1}$ is unbordered, and $|y_\delta| < |v| - |t|$.*

Proof. If $|y_\delta| \ge |v| - |t|$, then there is a suffix t_0 of $t'ut$ such that $t_0 b$ is a prefix of y_δ, and hence, a prefix of δ. This contradicts the maximality of δa w.r.t. $\lhd^a$ since $t_0 a$ is a suffix of $t'uta$, and hence, a suffix of δa. So, we have $|y_\delta| < |v| - |t|$.

In particular, we have that the δa-critical prefix of $\delta t^{-1}zuv$ is strictly longer than $\delta t^{-1}zut$, whence it can be written as $\delta t^{-1}zutv_\delta d$, where d is a letter. The definition of the critical prefix implies that $|tv_\delta d| \ge |v| - |y_\delta|$. Let g denote the δa-suffix of $\delta t^{-1}zutv_\delta$. Since δa does not occur in $t'uv$ by Claim 4.6, we have that $gd \neq \delta a$. Therefore gc is a prefix of δa and $c \neq d$. Moreover, we deduce $d \lhd^a c$ from Claim 4.6.

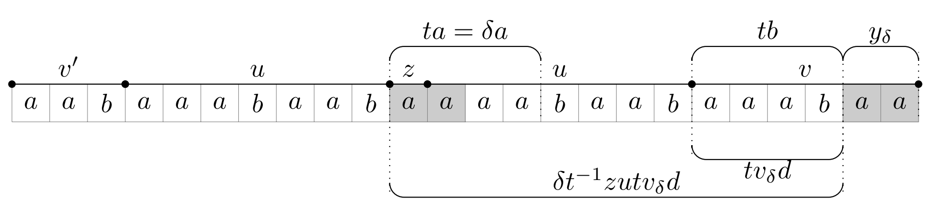
Suppose that $|tv_\delta d| > |v| - |y_\delta|$. Then there is a suffix g' of g such that $g'd$ is a prefix of y_δ, and hence, also of δ. We obtain a contradiction with the maximality of δa, since $g'c$ is a factor of δa. The situation is illustrated in the following figure.



Therefore $|tv_\delta d| = |v| - |y_\delta|$ and $\delta t^{-1}zuvy_\delta^{-1}$ is the δa-critical prefix of $\delta t^{-1}zuv$.

Suppose that $\delta t^{-1}zuvy_\delta^{-1}$ is bordered, and let h be its shortest border. The definition of the critical prefix implies that the δa-period of $\delta t^{-1}zuvy_\delta^{-1}$ is $|\delta t^{-1}zuvy_\delta^{-1}|$, whence $\delta a \leq_{\mathrm{p}} h$. Since $|h| < |u|$, we have that δa occurs in uv contradicting Claim 4.6.

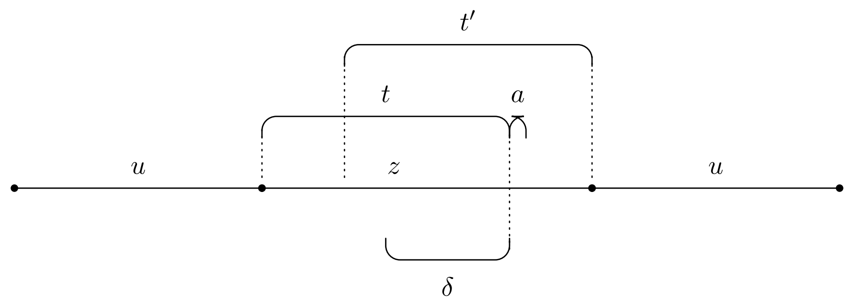
Our running example gives the following setting, with $d = b$.



Claim 4.8. *The word δ satisfies*

$$|\delta| > |t| + |t'| - |z|\,. \tag{4}$$

Proof. Suppose the contrary. Then δ lies within the overlap of ut and $t'u$ in uzu, as illustrated by the following figure.



This contradicts the maximality of δa since it occurs now twice in $t'uta$; see also Remark 2.2.

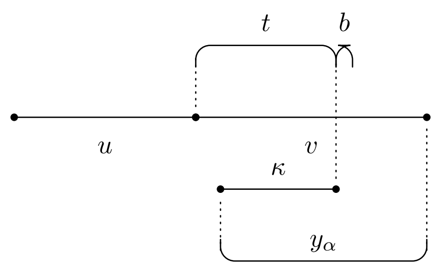
Remark 4.9. Similarly to the symmetric version of Claim 4.4, see Remark 4.5, we have a symmetric setting for Claim 4.7, too.

Provided that $t' \neq v'$, let $a't' \leq_{\mathrm{s}} v'uz$ and $b't' \leq_{\mathrm{s}} v'$ with $a' \neq b'$. Let δ' be the mirror analogue to δ. If $|v'| \leq |y't'|$, then Claim 4.7 translates to $y_\delta'^{-1} v'uzt'^{-1}\delta'$ is unbordered and $|y'_\delta| < |v'| - |t'|$.

Claim 4.4 is formulated for an arbitrary order $\lhd$. Since we want to combine results of § 4.1 with the present section, we shall identify $\lhd$ and $\lhd^a$. In particular, we have $b \lhd a$, and δa is the $\lhd$-maximum suffix of $t'uta$.

The conditions $|v| \leq |ty|$ and the $t \neq v$ now imply that utb and y_α have in uv an overlap κb. In other words, κ is a suffix of ut such that $uv = ut\kappa^{-1}y_\alpha$. Since $|y_\alpha| < |\alpha|$, we have

$$|t| > |v| - |\alpha| + |\kappa|\,. \tag{5}$$



Note that κb is a prefix of y_α, and κa a suffix of uta. We have the following claim.

Claim 4.10. *If t' is as short as possible, then*

$$|\kappa| > |t| + |t'| - |z| . \tag{6}$$

Proof. Similarly as above for δ, we deduce that κa cannot be a factor of the overlap of t and t' in z, otherwise α is not the $\lhd$-maximum suffix of $t'u$, a contradiction with Claim 4.2 on page 9.

4.3. Implied Inequalities

In this subsection, we summarize the properties proved in previous sections and conclude the proof of the main Theorem 4.1. We proceed by case distinction, which is based on whether or not $t = v$ ($t' = v'$); in addition to the main criterion of previous sections, that is, whether or not $|v| > |ty|$ ($|v'| > |t'y'|$).

Case 1: $t \neq v$ or $t' \neq v'$, but not both.

By symmetry, we assume $t \neq v$ and $t' = v'$ in the following. We also assume that t' is as short as possible. Note that this assumption does not change the situation, that is, we still have $t \neq v$ and $t' = v'$ (see (3)).

Subcase 1.1: $|v| > |ty|$

Claim 4.4 on page 11 yields $\tau(w) \geq |\gamma zuvy^{-1}|$. If $|v'| \leq |v|$, then the inequality $|\gamma| > |y|$ implies $\tau(w) > |zuv| \geq \frac{1}{2}|w|$; a contradiction to our assumption. We therefore have $|v'| > |v|$.

Claim 4.2 implies

$$|\gamma z| > |v'| \,. \tag{7}$$

Indeed, if $|\gamma z| \leq |v'|$, then $\gamma z \leq_{\mathrm{s}} t' = v'$, and hence, there is a maximum suffix ϑ of $t'u$ strictly longer than u contradicting Claim 4.2 (where we let ϑ be the maximum suffix of $t'u$ with respect to the same order to which γ is the maximum suffix of u).

We deduce a contradiction since now $\tau(w) \geq |\gamma zuvy^{-1}| > \frac{1}{2}|w|$ by

$$\begin{aligned}
2(|\gamma| + |z| + |u| + |v| - |y|) &> |v'| + |\gamma| + |z| + 2(|u| + |v| - |y|) && \text{(by (7))}\\
&> |v'| + |z| + 2(|u| + |v|) - |y| && \text{(by } |\gamma| > |y|)\\
&> |v'| + |z| + 2|u| + |v| && \text{(by } |v| > |y|)\\
&= |w|.
\end{aligned}$$

Subcase 1.2: $|v| \leq |ty|$

We obtain a contradiction by establishing a set of inequalities that do not have a common solution. Inequality (4) can be transformed into

$$\mathtt{L}_1 := |\delta| - |t| - |t'| + |z| - 1 \geq 0 \,.$$

Claim 4.7 on page 14 yields $|\delta zu| + 1 \leq \frac{3}{7}|w|$ which together with $|w| = |v| + |v'| + 2|u| + |z|$ gives

$$\mathtt{L}_2 := 3|v'| + 3|v| - |u| - 4|z| - 7|\delta| - 7 \geq 0 \,.$$

Moreover, since $|y_\delta| \leq |\delta|$, Claim 4.7 yields $|t^{-1}zuv| \leq \frac{3}{7}|w|$ and we obtain

$$\mathtt{L}_3 := 7|t| + 3|v'| - 4|v| - 4|z| - |u| \geq 0 \,. \tag{8}$$

The desired contradiction now follows from

$$21\,\mathtt{L}_1 + 4\,\mathtt{L}_2 + 3\,\mathtt{L}_3 = -7|uz\delta + 7| \,,$$

which is obtained keeping in mind that $t' = v'$.

Case 2: $t \neq v$ and $t' \neq v'$.

By symmetry, we can suppose $|v'| \leq |v|$, which implies $\tau(w) < |\gamma z u v y^{-1}|$, see the beginning of Subcase 1.1. Claim 4.4 now yields $|v| \leq |ty|$. As above in Subcase 1.2, we obtain $\mathtt{L}_1$, $\mathtt{L}_2$, $\mathtt{L}_3 \geq 0$. We need some more inequalities in this case for we assume $t' \neq v'$. Inequality (5) can be transformed into

$$\mathtt{L}_4 := |t| + |\alpha| - |\kappa| - |v| - 1 \geq 0\,,$$

and the inequality (6) into

$$\mathtt{L}_5 := |\kappa| - |t| - |t'| + |z| - 1 \geq 0\,.$$

We now exploit Remark 4.5 on page 13. If $\tau(w) \geq |{y'}^{-1} v' u z \gamma'|$, then using $|y'| < |\gamma'|$ and $\tau(w) \leq \frac{3}{7}|w|$ we obtain the inequality

$$\mathtt{L}_6 := 3|v| - 4|v'| - 4|z| - |u| - 7 \geq 0\,.$$

If, on the other hand, the inequality $|v'| \leq |t'y'|$ holds, then we can use Remark 4.9 and derive the mirror variant of (8), namely, the inequality

$$\mathtt{L}_6' := 7|t'| + 3|v| - 4|v'| - 4|z| - |u| \geq 0\,.$$

We now get

$$14\,\mathtt{L}_1 + 2\,\mathtt{L}_2 + 2\,\mathtt{L}_3 + 7\,\mathtt{L}_4 + 7\,\mathtt{L}_5 + 3\,\mathtt{L}_6' =$$
$$14\,\mathtt{L}_1 + 2\,\mathtt{L}_2 + 2\,\mathtt{L}_3 + 7\,\mathtt{L}_4 + 7\,\mathtt{L}_5 + 3\,\mathtt{L}_6 + 21(|t'| + 1) = -42 - 7|zu\alpha^{-1}|\,;$$

a contradiction again.

Case 3: $t = v$ and $t' = v'$.

This is the only case, in which we prove $\tau(w) = \pi(w)$, instead of a contradiction.

We have $\pi(w) \leq |uz|$ and, clearly, we can suppose that $\pi(w) > |u|$, since otherwise $\pi(w) = \tau(w) = |u|$. Let rs be a critical factorization of u. Then szr is unbordered of length $\pi(w)$, unless r is a prefix, and s is a suffix of z; see

Remark 2.5 on page 7. Suppose the latter possibility. Now, either one of the words uz and zu is unbordered of length $\pi(w)$ or u is both prefix and suffix of z. We are therefore left with the case $w = v'u^i z'u^j v$, with $i, j \geq 2$, where u is not a suffix of uz' and not a prefix of $z'u$. Note that z' cannot be empty. Moreover, v' is a suffix of u and v is a prefix of u, which implies

$$|v'| < |u|, \qquad |v| < |u| \tag{9}$$

by the maximality of z.

Suppose, without loss of generality, $i \leq j$. Similarly as above, we have that either $sz'u^{j-1}r$ or $z'u^j$ is unbordered. From $|u^j z'| \leq \frac{3}{7}|w|$ and from $|z'| > 0$ we deduce

$$|v'v| > \left(\frac{4}{3}j - i\right)|u|\,.$$

If $i < j$, then we obtain from $j \geq 3$ that $|v'v| > 2|u|$; a contradiction with (9). Therefore $i = j$.

If v' is a suffix of uz' and v a prefix of $z'u$, then we have $\pi(w) = \tau(w) = |z'u^j|$. Otherwise we obtain from Case 1 and Case 2 an unbordered factor of $v'uz'uv$ of length at least $\frac{3}{7}|v'uz'uv|$. Moreover, this factor contains u as a factor, which can be substituted with u^j to obtain an unbordered factor of w of length at least $\frac{3}{7}|v'u^j z'u^j v|$. Take, as an example, the word $\beta zuv_0 d$ from the proof of Claim 4.4 on page 11. If $\beta z'uv_0 d$ is unbordered, then clearly also the word $\beta z'u^j v_0 d$ is unbordered, which is a factor of w, and has the required length. It is not difficult to check that this happens with all words for which the proof used the main assumption that they are shorter than $\frac{3}{7}|w|$.

This concludes the proof of Theorem 4.1.

5. Conclusions

The relation between the period $\pi(w)$ of a word w and the length $\tau(w)$ of its longest unbordered factors has been investigated in this paper. Clearly, $\tau(w) \leq \pi(w)$. It is also not difficult to see that $\tau(w) = \pi(w)$ holds for long words, that is, for words, which are much longer than both $\tau(w)$ and $\pi(w)$. The question of interest is: When exactly is a word long enough so that $\tau(w) = \pi(w)$ is enforced? When the word length is expressed w.r.t. $\pi(w)$, it is well-known that

$$|w| > 2\pi(w) - 2 \qquad \text{implies} \qquad \tau(w) = \pi(w)\,.$$

Theorem 4.1 of the present paper makes the complementary statement

$$|w| \geq \frac{7}{3}\tau(w) \qquad \text{implies} \qquad \tau(w) = \pi(w)\,.$$

This solves a problem raised first by Ehrenfeucht and Silberger in 1979.

The bounds $2\tau(w)$ (see [1]) and $3\tau(w)$ (see [2]) have been previously conjectured, and several attempts in proving the latter have been made; see [3, 4, 5, 6, 9]. However, the bound proved above is (asymptotically) tight as demonstrated by an example in [2] with words of length $\frac{7}{3}\tau(w) - 4$ and $\tau(w) < \pi(w)$. For the sake of clarity we did not try to make the additive constant optimal in this paper. We only note that our arguments can be easily modified to obtain that already $|w| > \frac{7}{3}\tau(w) - \frac{8}{3}$ implies $\tau(w) = \pi(w)$. We do not consider this value of the additive constant to be too interesting since we conjecture that the example by Assous and Pouzet is optimal, that is

$$|w| > \frac{7}{3}\tau(w) - 4 \qquad \text{implies} \qquad \tau(w) = \pi(w)\,,$$

and, moreover, if $|w| = \frac{7}{3}\tau(w) - 4$ and $\tau(w) \neq \pi(w)$, then w is of the form given by (1).

Apart from the actual result, we would like to point out the proof techniques used to solve the Ehrenfeucht–Silberger problem. In particular, the notion of α-critical prefix of a word w (Definition 3.3) is used to find long unbordered factors in words with a large period, that is, words that do not have much of a global structure. We are confident that the investigation of α-critical prefixes of a word will lead to more insights in its structure, for example w.r.t. its local periods.

References

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

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

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

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

[5] J.-P. Duval, T. Harju, D. Nowotka, Unbordered factors and Lyndon words, Discrete Math. 308 (11) (2008) 2261–2264, (submitted in 2002).

[6] T. Harju, D. Nowotka, Minimal Duval extensions, Internat. J. Found. Comput. Sci. 15 (2) (2004) 349–354.

[7] T. Harju, D. Nowotka, Periodicity and unbordered words, in: STACS 2004 (Montpellier), Vol. 2996 of Lecture Notes in Comput. Sci., Springer-Verlag, Berlin, 2004, pp. 294–304.

[8] T. Harju, D. Nowotka, Periodicity and unbordered words: A proof of the extended Duval conjecture, J. ACM 54 (4).

[9] Š. Holub, A proof of the extended Duval's conjecture, Theoret. Comput. Sci. 339 (1) (2005) 61–67.

[10] M.-P. Schützenberger, A property of finitely generated submonoids of free monoids, in: Algebraic theory of semigroups (Proc. Sixth Algebraic Conf., Szeged, 1976), Vol. 20 of Colloq. Math. Soc. János Bolyai, North-Holland, Amsterdam, 1979, pp. 545–576.

[11] Y. Césari, M. Vincent, Une caractérisation des mots périodiques, C. R. Acad. Sci. Paris Sér. A 286 (1978) 1175–1177.

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

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