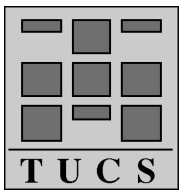
Minimal Duval Extensions

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 520
April 2003

ISBN 952-12-1149-0
ISSN 1239-1891

Abstract

A word $v = wu$ is a (nontrivial) Duval extension of the unbordered word w, if (u is not a prefix of v and) w is an unbordered factor of v of maximum length. After a short survey of the research topic related to Duval extensions, we show that, if wu is a minimal Duval extension, then u is a factor of w. We also show that finite, unbordered factors of Sturmian words are Lyndon words.

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

TUCS Laboratory
Discrete Mathematics for Information Technology

1 Introduction

In this paper a short survey about the research on the relationship between the length of a word and its unbordered factors is given. This line of research was introduced by Ehrenfeucht and Silberger [3] and Assous and Pouzet [1] in 1979. It was carried further and culminated in a strong conjecture by Duval [2] in 1982.

We will give a historical overview on this line of research, its main results and conjectures so far, in Section 2. This will lead to the concept of Duval extensions which are introduced in Section 3. In that section we will also show that unbordered, finite factors of Sturmian words are Lyndon words which we are not aware of being shown in the literature so far. Our main contribution is presented in Section 4. It shows a remarkable property of minimal Duval extensions.

We shall now introduce the main notations of this paper. We refer the reader to [5, 6] for more basic and general definitions.

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 $w \in A^*$. Then we can express w as a sequence of letters $w_{(1)}w_{(2)}\cdots w_{(n)}$ where $w_{(i)} \in A$ is a letter, for every $1 \leq i \leq n$. We denote the length n of w by $|w|$. Note, that $|\varepsilon| = 0$. A word w is called *primitive* if it cannot be factored such that $w = u^k$ for some $k \geq 2$. Let $w = uv$ for some words u and v. Then vu is called *conjugate* of w. Let $[w]$ denote the set of all conjugates of w. Note, that $w \in [w]$.

A nonempty word u is called a *border* of a word w, if $w = uv = v'u$ for some 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. 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 $\lhd_A$ be an ordering of $A = \{a_1, a_2, \ldots, a_n\}$, say $a_1 \lhd_A a_2 \lhd_A \cdots \lhd_A a_n$. Then $\lhd_A$ induces a *lexicographic order*, also denoted by $\lhd_A$, on A^* such that

$$u \lhd_A v \iff u \leq v \quad \text{or} \quad u = xau' \text{ and } v = xbu' \text{ with } a \lhd_A b$$

where $a, b \in A$. We write $\lhd$ for $\lhd_A$, for some alphabet A, if the context is clear.

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 u and v are primitive, but w is not. Furthermore, $[u] = \{aaab, aaba, abaa, baaa\}$ is the set of all conjugates of u. Let $a \lhd b$. Then $u \lhd w \lhd v$. We have that a is

the shortest border of u and w, whereas ba is the shortest border of v. The smallest unbordered factor of w has length three.

2 On the Maximum Length of Unbordered Factors

When the length of unbordered factors of a word is investigated, that is usually done in terms of the length of the word and its minimum period.

Lets make our terminology more precise. Consider a word w over some alphabet A. 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 $\partial(w)$. Let $\mu(w)$ denote the maximum length of unbordered factors of w. For example, let $w = abaabbaaba$, then $\partial(w) = 7$ and $\mu(w) = 6$.

Clearly, the maximum length of unbordered factors $\mu(w)$ of w is bound by the period $\partial(w)$ of w. We have

$$\mu(w) \leq \partial(w)$$

since for every factor v of w, with $\partial(w) < |v|$, the prefix $v_{(1)}v_{(2)}\cdots v_{(|v|-\partial(w))}$ of v is also a suffix of v by the definition of period.

It is a natural question to ask at what length of w is $\mu(w)$ necessarily maximal, that is, $\mu(w) = \partial(w)$. Of course, the length of w is considered with respect to either $\mu(w)$ or $\partial(w)$.

In 1979 Ehrenfeucht and Silberger [3], as well as, Assous and Pouzet [1] addressed this question first. Ehrenfeucht and Silberger [3] stated

Theorem 1. *If* $2\partial(w) \leq |w|$ *then* $\mu(w) = \partial(w)$.

They also established that every primitive word w has at least σ-many unbordered conjugates, where σ is the number of different letters occuring in w, which leads directly to

Theorem 2. *If* $2\partial(w) - \sigma \leq |w|$ *then* $\mu(w) = \partial(w)$.

However, this result was stated by Duval [2] only in 1981.

The real challenge, though, turned out to be giving a bound on the length of w with respect to $\mu(w)$. It was conjectured in [3] that $2\mu(w) \leq |w|$ implies $\mu(w) = \partial(w)$. However, Assous and Pouzet gave the following counter example. Let

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

for which $|w| = 7n + 10$ and $\mu(w) = 3n + 6$ and $\partial(w) = 4n + 7$ contradicting that conjecture. They themselves gave the following conjecture.

Conjecture 3. *Let $f\colon \mathbb{N} \to \mathbb{N}$ such that $f(\mu(w)) \leq |w|$ implies $\mu(w) = \partial(w)$. Then*

$$f(\mu(w)) \leq 3\mu(w) \ .$$

In 1982 Duval [2] established the following.

Theorem 4. *If $4\mu(w) - 6 \leq |w|$ then $\mu(w) = \partial(w)$.*

He also stated Conjecture 6 (see next section) about what was later called Duval extensions that would imply

$$\text{If} \quad 3\mu(w) \leq |w| \quad \text{then} \quad \mu(w) = \partial(w) \ .$$

3 Duval Extensions

In the previous section we recalled a question initially raised by Ehrenfeucht and Silberger [3]. The problem was to estimate a bound on the length of w, depending on $\mu(w)$, such that $\mu(w) = \partial(w)$. Duval [2] introduced a restricted version of that problem by assuming that w has an unbordered prefix of length $\mu(w)$. However, lets first fix 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 is called *trivial*, if $\partial(wu) = \mu(wu)$. 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 5. *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.

In 1982 Duval [2] stated the following conjecture.

Conjecture 6. *Let wu be a nontrivial Duval extension of w. Then $|u| < |w|$.*

It follows directly from this conjecture that for any word w, we have that $3\mu(w) \leq |w|$ implies $\mu(w) = \partial(w)$. This conjecture has remained popular throughout the years, see for example Chapter 8 in [6].

Duval extensions have also become a subject of interest on their own. In particular the set of words having no nontrivial Duval extension has been investigated in [4] and [8].

Infinite words of minimal subword complexity are called *Sturmian* words, cf. [9, 6]. Let us consider finite factors of Sturmian words in the following, and lets simply call them Sturmian words. Mignosi and Zamboni showed the following uniqueness result for Duval extensions in [8].

Theorem 7. *Unbordered Sturmian words have no nontrivial Duval extension.*

This result was improved by the authors of this paper in [4] to Lyndon words. Let a primitive word w be called *Lyndon* word if it is minimal among its conjugates, that is, if $w \lhd v$ for every $v \in [w]$ and some arbitrary order $\lhd$ on A, cf. [7, 6]. Note, that Lyndon words are unbordered.

Theorem 8. *Lyndon words have no nontrivial Duval extension.*

Theorem 10 states that unbordered Sturmian words are indeed Lyndon words. The following lemma will be used to prove that result.

Let $\tau\colon A^* \to B^*$ be a morphism, and $\lhd_A$ and $\lhd_B$ be orders on A and B, respectively, such that

$$a_1 \lhd_A a_2 \implies \tau(a_1) \lhd_B \tau(a_2) \tag{1}$$

for every $a_1, a_2 \in A$, and $\tau(a)$ is a Lyndon word w.r.t. $\lhd_B$ for every $a \in A$.

Lemma 9. *If $w \in A^*$ is a Lyndon word, then $\tau(w)$ is a Lyndon word.*

Proof. Let $|w| = n$. Assume w is a Lyndon word and $\tau(w)$ is not a Lyndon word. Therefore, $\tau(w) = xy$ such that yx is minimal w.r.t. $\lhd_B$, and x and y are not empty.

If $x = \tau\big(w_{(1)}w_{(2)}\cdots w_{(i)}\big)$ and $y = \tau\big(w_{(i+1)}w_{(i+2)}\cdots w_{(n)}\big)$ with $1 \leq i < n$, then we have an immediate contradiction by (1).

Therefore, there exists an i, where $1 \leq i \leq n$, and $\tau(w_{(i)}) = v_1v_2$ such that we have $x = \tau\big(w_{(1)}w_{(2)}\cdots w_{(i-1)}\big)v_1$ and $y = v_2\tau\big(w_{(i+1)}w_{(i+2)}\cdots w_{(n)}\big)$ and $v_1, v_2 \neq \varepsilon$. That implies $v_2 \lhd_B v_1v_2$, and we have $v_1 = u^j$ and $v_2 = u^k$,

for some primitive u and $j, k \geq 1$, since $v_1 v_2$ is a Lyndon word by assumption. But now, either

$$v_1 yx v_1^{-1} \lhd_B yx \qquad \text{or} \qquad v_2^{-1} yx v_2 \lhd_B yx \ ,$$

a contradiction. The following theorem shows that Theorem 8 implies Theorem 7.

Theorem 10. *Every unbordered Sturmian word is a Lyndon word.*

Proof. Let $u \in \{a, b\}$ be an unbordered Sturmian word. Assume u begins with a and ends with b without restriction of generality. The case is clear if $u = ab^k$ for some $k \geq 1$. Assume a occurs at least twice in u. Then $u = ab^k vab^{k+1}$ and u can be factored into ab^k and ab^{k+1} for some $k \geq 1$. Let $\tau \colon \{a, b\}^* \to \{a, b\}^*$ such that $\tau(a) = ab^k$ and $\tau(b) = ab^{k+1}$. Now, let $w = \tau(u)$ and we have that w is an unbordered Sturmian word that begins with a and ends in b. By induction w is a Lyndon word w.r.t. $a \lhd b$ and u is a Lyndon word w.r.t. $\lhd$ by Lemma 9.

The converse of Theorem 10 is certainly not true. Indeed, consider the word *aabbab* which is a Lyndon word but not a Sturmian word since it contains four factors of length two.

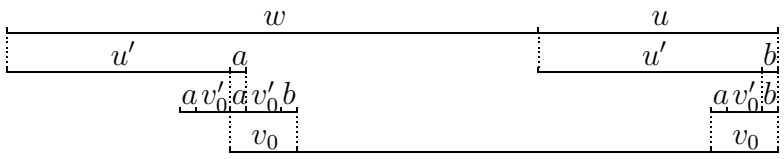
Another property of Duval extensions will be introduced in the next section.

4 Minimal Duval Extensions

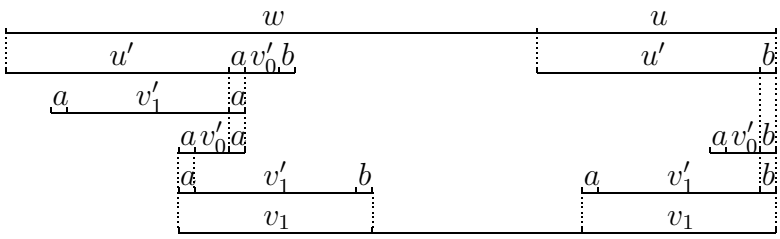
The minimal Duval extension of a word w is the smallest prefix of a nontrivial Duval extension of w such that the prefix itself is a nontrivial Duval extension of w. The following theorem gives a rather surprising property of nontrivial Duval extensions.

Theorem 11. *Let wu be a minimal Duval extension of w. Then u is a factor of w.*

Proof. By assumption, we have $w = u'aw'$ and $u = u'b$ where $a, b \in A$ and $a \neq b$. Consider the shortest border v_0 of $aw'u'b$ which exists by assumption since $|aw'u'b| = |w| + 1$. If $|v_0| \geq |u|$, then the claim holds. Thus, assume that $|v_0| < |u|$.

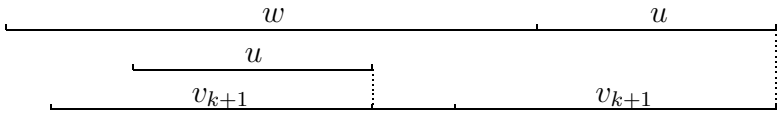


Let $v_0 = av_0'b$. Then $v_0'b \leq w'u$ and $av_0' \preccurlyeq u'$. The factor $av_0'aw'u$ has a shortest border v_1 such that $|av_0'av_0'| < |v_1|$ because otherwise v_0 is not the shortest border of $aw'u$. Again, if $|v_1| \geq |u|$, then the claim follows. Assume that $|v_1| < |u|$.



Let $v_1 = av_1'b$. Then $v_1'b \leq w'u$ and $av_1' \preccurlyeq u'$. The factor $av_1'aw'u$ has a shortest border v_2 such that $|av_1'av_1'| < |v_2|$ because otherwise v_1 is not the shortest border of $av_0'aw'u$.

In this way, we get a sequence of suffixes $v_0, v_1, \ldots, v_k$ of wu growing in length such that $|v_0| < \cdots < |v_k|$, and moreover v_i occurs in w for every $0 \leq i \leq k$.



If $|v_k| < |u|$, this sequence of suffixes of wu can be continued. Since wu is of finite length, the sequence is finite. Assume thus that k is the maximum index for which $|v_k| < |u|$. Then $|u| \leq |v_{k+1}|$, and hence $u \preccurlyeq v_{k+1}$, which proves the claim.

Acknowledgements

The authors would like to thank Julien Cassaigne for pointing out that unbordered Sturmian words might be Lyndon words which led us to the effort of proving it.

References

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

[2] 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.

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

[4] 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.

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

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

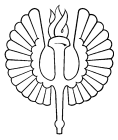
[7] R. C. Lyndon. On Burnside's problem. *Trans. Amer. Math. Soc.*, 77:202–215, 1954.

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

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

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**