Periodicity and Unbordered Segments of Words

Tero Harju Dirk Nowotka

Turku Centre for Computer Science (TUCS)
Department of Mathematics
University of Turku, Finland

We shall give an introduction to the problem area concerning the well known Duval's conjecture, which was announced to be solved in [12].

1 Introduction

In 1979 Ehrenfeucht and Silberger published an article [9] where the relationship between the length of a word and the maximum length of its unbordered factors (segments) was for the first time investigated. Periodicity and borderedness are two basic properties of words that play a rôle in many areas of computer science such as string searching algorithms [13, 3, 7], data compression [20, 6], and codes [2], which are classical examples, but also computational biology, e.g., sequence assembly [17] or superstrings [4], and serial data communications systems [5]. It is well known that these two word properties do not exist independently from each other. However, no clear relationship has been established so far, despite substantial recent progress in that area. We chose the title of Ehrenfeucht and Silberger's paper for our essay to underline that this paper illustrates the way this line of research has evolved in the last 24 years.

In Section 2 we give a historical overview on this line of research, its main results and conjectures so far. This will lead to the concept of Duval extensions which are introduced in Section 3. We conclude with Section 4.

First, we shall introduce the main notations of this paper. We refer the reader to [14, 15] 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 $1 \leq i \leq n$. Let $1 \leq i \leq j \leq n$, then $w_{(i)}w_{(i+1)}\cdots w_{(j)}$is called a factor or segment of w. 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.

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*. Suppose $w = uv$, then u is called a *prefix* of w denoted by $u \leq w$.

Let $\lhd$ be an ordering of $A = \{a_1, a_2, \ldots, a_n\}$, say $a_1 \lhd a_2 \lhd \cdots \lhd a_n$. Then $\lhd$ induces a *lexicographic order*, also denoted by $\lhd$, on A^* such that

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

where $a, b \in A$.

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, $\{aaab, aaba, abaa, baaa\}$ is the set of all conjugates of u. Let $a \lhd b$. Then $u \lhd w \lhd v$. The largest unbordered factor of w has length three.

2 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 of length n 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 [9], as well as, Assous and Pouzet [1] addressed this question first. Ehrenfeucht and Silberger [9] 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 [8] only in 1982.

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 [9] that $2\mu(w) \leq |w|$ implies $\mu(w) = \partial(w)$. However, Assous and Pouzet [1] gave the following counter example.

Example 3. *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 $\partial(w) = 4n + 7$ *and* $\mu(w) = 3n + 6$ *and* $|w| = 7n + 10 = 7/3\mu(w) - 4$.

Assous and Pouzet also gave the following conjecture.

Conjecture 4. *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 [8] established the following result.

Theorem 5. *If* $4\mu(w) - 6 \leq |w|$ *then* $\mu(w) = \partial(w)$.

He also stated Conjecture 7 (see Duval's conjecture in the following section) about what was later called Duval extensions, a conjecture that implies

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

3 Duval Extensions

In the previous section we recalled a question initially raised by Ehrenfeucht and Silberger [9]. The problem was to estimate a bound on the length of w, depending on $\mu(w)$, such that $\mu(w) = \partial(w)$. Duval [8] introduced a restricted version of that problem by assuming that w has an unbordered prefix of length $\mu(w)$. Let us fix some more notations first.

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 6. *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 [8] stated the following conjecture.

Conjecture 7 (Duval). *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 remained popular throughout the years, see for example Chapter 8 in [15]. However, recently the authors of this essay established [12] the following result which implies Duval's conjecture.

Theorem 8. *Let wu be a nontrivial Duval extension of w. Then $|u| < |w| - 1$.*

We have the following corollary.

Corollary 9. *If $3\mu(w) - 2 \leq |w|$ then $\mu(w) = \partial(w)$.*

This corollary gives the best bound on the general case so far. However, this result does not give a final answer to the question about the relation between $|w|$ and $\mu(w)$. Since the best example known to us is Example 3 in the previous section which shows that there is an arbitrary long word w such that $|w| = 7/3\mu(w) - 4$ and $\mu(w) < \partial(w)$. In general the precise bound is still unknown.

Nevertheless, the bound of Theorem 8 is tight as the following example shows.

Example 10. *Let $w = a^iba^{i+j}bb$ and $u = a^{i+j}ba^i$ where $i, j \geq 1$. It is easy to check that*

$$w.u = a^iba^{i+j}bb.a^{i+j}ba^i$$

is a nontrivial Duval extension of w of length $|w| - 2$.

Further knowledge about structural properties of Duval extension such as the following conjecture would certainly help to estimate the precise bound in the general case. Let us call a nontrivial Duval extension wu of w *maximal* if $|u| = |w| - 2$.

Conjecture 11. *Let $w = w'ab^k$ for some $k \geq 1$. If wu is a maximal Duval extension of w, then b^k does not occur in w'.*

A further property of Duval extensions was established in [11]. Recall, that a 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 another property of nontrivial Duval extensions.

Theorem 12. *Let wu be a minimal Duval extension of w. Then u is a factor of w.*

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 [10] and [18]. The first attempt to characterize the set of words

that have no nontrivial Duval extension was done investigating well-known sets of words like unbordered finite factors of Sturmian words and Lyndon words.

Infinite words of minimal subword complexity are called *Sturmian* words, cf. [19, 15]. Minimal subword complexity means that a Sturmian word contains exactly $n+1$ different factors of length n for every $n \geq 1$. 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 [18].

Theorem 13. *Unbordered Sturmian words have no nontrivial Duval extension.*

This result was improved by the authors of this paper in [10] 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 conjugate v of w and some arbitrary order $\lhd$ on A, cf. [16, 15]. Note, that Lyndon words are unbordered.

Theorem 14. *Lyndon words have no nontrivial Duval extension.*

The following theorem [11] shows that Theorem 14 implies Theorem 13.

Theorem 15. *Every unbordered Sturmian word is a Lyndon word.*

The converse of Theorem 15 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. A precise characterization of the set of words that have no nontrivial Duval extension is still unknown.

4 Conclusions

We have recalled the problem of estimating the relationship between the length of a word and the maximum length of its unbordered factors. The final answer is still unknown. However, quite some progress has been made since the problem was raised in 1979. In particular the special case of Duval extensions raised attention and led to new results. For example, the long standing conjecture by Duval was just recently solved. However, open problems remain about the structure of Duval extensions and words that have no nontrivial Duval extension. Further research on those questions are likely to lead to a final answer to the general case.

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] 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.

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

[8] 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.

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

[10] 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.

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

[12] T. Harju and D. Nowotka. Periodicity and unbordered words. technical report 523, Turku Centre of Computer Science (TUCS), Turku, Finland, April 2003. submitted.

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

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

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

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

[17] 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.

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

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

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