ELSEVIER

Theoretical Computer Science 218 (1999) 135–141

Theoretical
Computer Science

Partial words and a theorem of Fine and Wilf

Jean Berstel[a], Luc Boasson[b,*]

[a] *Institut Gaspard Monge (IGM), Université Marne-la-Vallée, 5, boulevard Descartes, 77 Marne-la-Vallée Cédex 2, France*

[b] *Laboratoire d'informatique algorithmique: fondements et applications (LIAFA) Université Denis-Diderot 2, place Jussieu, 75251 Paris Cédex 05, France*

Abstract

A partial word is a word that is a partial mapping into an alphabet. We prove a variant of Fine and Wilf's theorem for partial words, and give extensions of some general combinatorial properties of words.

Keywords: Combinatorial problems; Words; Formal languages

1. Introduction

A *partial word* of length n over an alphabet A is a partial function

$$w : \{0,\ldots,n-1\} \rightarrow A.$$

The *domain* of w is the set $D(w)$ of positions $p \in \{0,\ldots,n-1\}$ such that $w(p)$ is defined. The set $H(w) = \{0,\ldots,n-1\} \setminus D(w)$ is the set of *holes* of w. A usual word over an alphabet A is just a partial word without holes.

Partial words appear in comparing genes. Indeed, alignment of two sequences can be viewed as a construction of two partial words that are *compatible* in a sense that will be developed below.

The aim of this note is to examine to which extent some elementary combinatorial properties of words remain true for partial words. As we shall see, these properties still hold when words have one hole, but become false as soon as words have two holes.

* Corresponding author. E-mail: luc.boasson@liafa.jussieu.fr.

PII: S0304-3975(98)00255-2

2. Notation

To each partial word w over A we can associate a total word $w_\diamond$ (its *companion*) over an augmented alphabet $A_\diamond = A \cup \{\diamond\}$ by setting

$$w_\diamond(p) = \begin{cases} w(p) & \text{if } p \in D(w), \\ \diamond & \text{if } p \in H(w). \end{cases}$$

The mapping $w \mapsto w_\diamond$ is a bijection, so many relevant notions for words, such as concatenation, powers, and so on, may be transported to partial words this way.

Observe that the symbol $\diamond$ is not a "do not care" symbol, as for pattern matching, but rather a "do not know" symbol.

Example. The (total) word $w_\diamond = abc\diamond\diamond cd$ corresponds to a partial word w of length 7, with set of holes $H(w) = \{3,4\}$.

A partial word w has *period* p or is *p-periodic* if, for all $i, j \in D(w)$,

$$i \equiv j \pmod{p} \Rightarrow w(i) = w(j).$$

As an example, the word w with companion $w_\diamond = ab\diamond a\diamond cabc$ is 3-periodic. Observe that, despite the fact that the length of w is a multiple of the period, w is *not* a power of a shorter word. This shows a clear difference between partial and total (or full) words.

A partial words w is *locally p-periodic* if

$$i, i+p \in D(w) \Rightarrow w(i) = w(i+p).$$

A locally periodic total word is always periodic. This does not hold for partial words. As an example, the word $w = abc\diamond bcd$ is locally 3-periodic but is not 3-periodic unless $a = d$.

3. Fine and Wilf's theorem revisited

A well-known result due to Fine and Wilf [2] (see also the exposition in [1]) is the following:

Theorem 3.1. *If a (total) word x has periods p and q, and has length at least $p + q - \gcd(p,q)$, then x has also period $\gcd(p,q)$.*

In the particular case where the periods p and q are relatively prime, it suffices that $|x| \geqslant p + q - 1$ to get that x is 1-periodic.

We consider the case of partial words, and prove the following analog of Theorem 3.1.

Theorem 3.2. *Let w be a partial word of length n which is locally p-periodic and locally q-periodic. If $H(w)$ is a singleton and if $n \geqslant p+q$, then w is (strongly) $\gcd(p,q)$-periodic.*

In the special case where p and q are relatively prime, our result implies Fine and Wilf's theorem for relatively prime periods. Indeed, the latter is obtained in the case where the hole is just at the beginning or at the end of the word.

The bound for the length n is sharp. Indeed, the partial word of length 12

$$aaaabaaaa\diamond aa$$

has (strong) periods 5 and 8, has a unique hole but is not 1-periodic.

The result does not hold for two holes. Indeed, the partial word with companion

$$ab\diamond aba\diamond ba$$

of length 9 has periods 3 and 5 and two holes, without being 1-periodic. This is a special case of an infinite set. Consider

$$x_\diamond^{(m)} = (ab)^m \diamond (ab)^m a \diamond (ba)^m.$$

Each word $x^{(m)}$ has length $6m+3$, and is easily checked to have periods $2m+1$ and $2m+3$, without being 1-periodic.

4. Proof

The proof is along the lines of the proof given in [1] with one modification where it appears to be necessary. We will illustrate it by an example.

It suffices to prove the result for $n=p+q$, for, if $|w|>p+q$, it holds for every factor of w of length $p+q$, and thus also for w itself.

Assume first that the result holds for relatively prime periods p and q, and consider the case where $d=\gcd(p,q)>1$. Set $n=dn'$, and define d partial words

$$w^{(k)} = w(k)w(k+d)\cdots w(k+(n'-1)d), \quad k=0,\ldots,d-1.$$

Setting $p=p'd$, $q=q'd$, each of the $w^{(k)}$ is locally p'-periodic, and locally q'-periodic and has length $n'=p'+q'$. Consequently, it is 1-periodic by our assumption, and w is d-periodic.

It remains to prove the result for relatively prime p and q. We may assume $p<q$. Consider the function

$$f:\{0,\ldots,p-1\} \to \{0,\ldots,p-1\}$$

defined, for $k \in \{0,\ldots,p-1\}$, by: $f(k)$ is the unique integer in $\{0,\ldots,p-1\}$ that is congruent to $k+q$ modulo p.

Since $f(k) \equiv k+q \pmod p$, it follows that $f^2(k) \equiv f(k)+q \equiv k+2q \pmod p$ and in general $f^h(k) \equiv k+hq \pmod p$ for all $k \in \{0,\dots,p-1\}$ and $h \geqslant 0$. Thus $f^p(k)=k$ and $f^h(k) \neq f^{h'}(k)$ for $0 \leqslant h < h' < p$. Thus, for every $k \in \{0,\dots,p-1\}$, one has

$$\{k, f(k), f^2(k), \dots, f^{p-1}(k)\} = \{0,\dots,p-1\}.$$

Define now the *path* of $k \in \{0,\dots,p-1\}$ to be the subset of $\{0,\dots,p+q-1\}$ given by

$$P(k) = \{k, k+q, k+q-p, k+q-2p, \dots, f(k)\}.$$

It starts with a big (forward) step $k \mapsto k+q$, followed by a sequence of small (backward) steps of the form $\ell \mapsto \ell - p$ taking eventually the number back into $\{0,\dots,p-1\}$. The set $H(w)$ being a singleton, set $H(w)=\{r\}$.

Before proceeding with the proof, let us give an example.

Let $p=5$ and $q=17$, and thus $|w|=22$. There are 5 paths

$$P(0)=\{0,17,12,7,2\}, \qquad P(1)=\{1,18,13,8,3\},$$

$$P(2)=\{2,19,14,9,4\}, \qquad P(3)=\{3,20,15,10,5,0\}.$$

$$P(4)=\{4,21,16,11,6,1\},$$

So, the permutation f is given by

$$(0,2,4,1,3).$$

The proof that follows is divided into two cases, according to the value of the hole r:
- $0 \leqslant r \leqslant 4$, for instance $r=2$. Then since $f(2)=4$, we obtain $w(4)=w(1)=w(3)=w(0)=w(i)$ for $i \geqslant 5$, as required.
- $5 \leqslant r \leqslant 21$, for instance $r=14$. Since 14 is in the path $P(2)$, all the other paths together ensure again that all letters are equal.

We now proceed with the formal proof.

If r is not in $P(k)$, then $w(k)=w(f(k))$, because weak q-periodicity implies $w(k)=w(k+q)$, and weak p-periodicity implies $w(k+q)=w(k+q-p)=\dots=w(f(k))$.

The sets $P(k) \setminus \{k\}$ are pairwise disjoint (and in fact are a partition of $\{0,\dots,p+q-1\}$). This means that the hole r belongs to exactly one of the sets $P(k) \setminus \{k\}$, and r belongs to at most two $P(k)$. More precisely:
- If $0 \leqslant r \leqslant p-1$, then one has $w(f(r))=w(f^2(r))=\dots=w(f^{p-1}(r))$. The numbers $f(r),\dots,f^{p-1}c(r)$ are exactly the integers $\{0,\dots,p-1\} \setminus \{r\}$, and this means precisely that the partial word $w(0)\cdots w(p-1)$ is 1-periodic. It follows easily that w is 1-periodic.
- If $p \leqslant r \leqslant p+q-1$, then r is in $P(t)$, for t such that $t+q \equiv r \pmod p$. One has $w(f(t))=w(f^2(t))=\dots=w(f^p(t))$, and as above this means that the total word $w(0)\cdots w(p-1)$ is 1-periodic. Again, it follows easily that w is 1-periodic. □

5. Further results

In this section, we present some analogous of well-known elementary combinatorial properties of words, extended to partial words. It appears that several implications still hold for partial words with a single hole, but become false for words with two holes.

Given two partial words x and y of the same length, we say that x is contained in y or that y contains x, and we write $x \subset y$, if $D(x) \subset D(y)$ and $x(k) = y(k)$ for all k in $D(x)$. Two words x and y are *compatible* and we write $x \uparrow y$ if there exists a word z that contains both x and y. In this case, the smallest word containing x and y is denoted by $x \vee y$ and is defined by $D(x \vee y) = D(x) \cup D(y)$

We start with several straightforward rules for computing with partial words.

$$
\begin{array}{ll}
x \uparrow z,\ \ y \uparrow t \Rightarrow xy \uparrow zt & \text{(multiplication)},\\
xy \uparrow zt,\ \ |x| = |z| \Rightarrow x \uparrow z \text{ and } y \uparrow t & \text{(simplification)},\\
x \uparrow y,\ \ z \subset x \Rightarrow z \uparrow y & \text{(weakening)}.
\end{array}
$$

From these rules, we get easily the following.

Lemma 5.1 (Levi's lemma). *Let x, y, z, t be partial words. If $xy \uparrow zt$ and $|x| \leqslant |z|$, there exists a factorisation $z = ps$ such that $x \uparrow p$ and $y \uparrow st$.*

Proof. Set indeed $z = ps$ with $|x| = |p|$. Then $xy \uparrow pst$ and the simplification rule gives the result. □

Theorem 5.2. *Let x and y be partial words such that xy has at most one hole. The following are equivalent:*

1. $xy \uparrow yx$.
2. $x \subset z^n, y \subset z^m$ *for some word z and integers n, m.*
3. $x^k \uparrow y^\ell$ *for some integers k, ℓ.*

As the proof shows, equivalence (2) ⇔ (3) and implication (2) ⇒ (1) hold without any condition.

On the contrary, implication (1) ⇒ (2) is false even if $xy \vee yx$ has no hole, as shown by the example $x_\diamond = \diamond bb$, $y_\diamond = abb\diamond$.

We start with a lemma.

Lemma 5.3. *Let x be a partial word and let u and v be two (full) words. If x has only one hole and if $x \subset uv$ and $x \subset vu$, then $uv = vu$.*

Proof. We may suppose that $|u| \leqslant |v|$. Set $v = u'v'$, with $|u'| = |u|$, and set $x = yz$ with $|y| = |u|$. Here u', v' are full words, and y, z are partial words.

From $yz \subset uv$, we get $y \subset u$ and $z \subset v$. Similarly, since $yz \subset vu = u'v'u$, we get $y \subset u'$ and $z \subset v'u$. Two cases arise.

Case 1: y has no hole (and z has one hole). Then $y = u = u'$, and $z \subset v = uv'$, $z \subset v'u$. By induction, $uv' = v'u$ whence $uv = vu$.

Case 2: y has one hole (and z has no hole). Then $y \subset u$, $y \subset u'$, and $z = v = v'u = u'v'$. Thus u and v' are conjugate full words, and there exists two words s and t such that $u = st$, $u' = ts$ and $v' = (ts)^n t$ for some $n \geqslant 0$. Since $y \subset st$, $y \subset ts$, by induction $st = ts$, whence $uv = vu$. □

Proof of the theorem. (1)$\Rightarrow$(2). If x and y have same length, then the simplification rule shows that $x \uparrow y$. Since either y or x has no hole, this implies $x \subset y$ or vice-versa. Assume $|x| < |y|$. By Levi's lemma, there is a factorization $y = ut$, with $|u| = |x|$, such that

$$x \uparrow u \quad \text{and} \quad ut \uparrow tx. \tag{1}$$

We distinguish three cases.

(a) The hole is in x. Then $x \uparrow u$ implies $x \subset u$, whence $xt \subset ut$. By the weakening rule, the second relation in (1) gives $xt \uparrow tx$. Since $x \subset u$, this implies $xt \subset tu$. By the previous lemma, $ut = tu$, and there exists a word z such that $t = z^n$ and $u = z^m$, so that $y = z^{n+m}$ and $x \subset z^m$.

(b) The hole is in u. This case is symmetric to the previous one.

(c) The hole is in t. Then (1) gives $x = u$ and $xt \uparrow tx$. By induction, $x \subset z^n$ (whence $x = z^n$) and $t \subset z^m$. By the multiplication rule, $y = ut \subset z^{n+m}$.

(2)$\Rightarrow$(1). The multiplication rule gives $xy \uparrow z^{n+m}$ and $yx \uparrow z^{n+m}$. Thus $xy \uparrow yx$.

(2)$\Rightarrow$(3). By the multiplication rule, $x^m \subset z^{nm}$ and $y^n \subset z^{nm}$, showing that $x^m \uparrow y^n$.

(3)$\Rightarrow$(2). The proof is partly similar to that of Fine and Wilf's theorem. Clearly, the result holds if $D(x) = \emptyset$ or $D(y) = \emptyset$. Set $p = |x|$ and $q = |y|$, and suppose first that p and q are relatively prime. Then $x^q \uparrow y^p$ by the simplification rule.

Let $i \in D(x)$. The numbers $i, i+p, \ldots, i+(q-1)p$ belong to $D(x^q)$. All these letters in x^q are the same, say the letter a. The remainders modulo q of $i, i+p, \ldots, i+(q-1)p$ are precisely the numbers $0, 1, \ldots, q-1$ (not necessarily in this order). Indeed, it suffices to show that the remainders are all distinct, and assuming

$$i + \lambda p \equiv i + \mu p \pmod{q}$$

one gets that q divides $(\lambda - \mu)p$, and since $(p,q) = 1$, that q divides $\lambda - \mu$, whence $\lambda = \mu$.

It follows that $y \subset a^q$, whence also $x \subset a^p$.

Assume next that $(p,q) = d$, and set $p = p'd$, $q = q'd$. Then $x^{q'} \uparrow y^{p'}$ a,d $|x^{q'}| = |y^{p'}| = p'q'd$. Define, for $0 \leqslant h \leqslant d-1$,

$$x_h = x(h)x(h+d)\cdots x(h+(p'-1)d),$$
$$y_h = y(h)y(h+d)\cdots y(h+(q'-1)d).$$

Then $|x_h| = p'$, $|y_h| = q'$, and $x_h^{q'} \uparrow y_h^{p'}$. By the previous argument, one gets $x_h \subset a_h^{p'}$ and $y_h \subset a_h^{q'}$ for some letter h, whence $x \subset (a_0 \cdots a_{d-1})^{p'}$ and $y \subset (a_0 \cdots a_{d-1})^{q'}$. □

It is natural to look for a similar result concerning conjugacy. Consider nonempty partial words x, y, t such that $xy \uparrow yt$. It is reasonable to conjecture that, under mild assumptions, there exist partial words u, v such that

$$x \subset uv, \quad t \subset vu, \quad y \subset (uv)^n u.$$

However this is false, even if xyt has a single hole. Consider indeed $x = a$, $y = \diamond bb$ and $t = b$. Then $xy = a\diamond bb$ and $yt = \diamond bbb$ are clearly compatible, and there are no words u, v such that $a = uv$ and $b = vu$.

6. Concluding remarks

The order $x \subset y$ on partial words introduced in the previous section is well known as the "less defined" order in denotational semantics. It is simply defined by taking $\diamond$ as the bottom element in the flat order over the alphabet, and extending this order to sequences. It might therefore be interesting (as suggested by Olivier Carton) to consider words over ordered alphabets (not necessarily linear orders) and to extend other combinatorial results.

Combinatorics of partial words will presumably not be as rich as the combinatorics of words. It is clear that the more holes are in words, the more degrees of freedom exist to combine and to compare them. It is somewhat astonishing that if only one hole is allowed, several classical results still hold, however they become false as soon as there are two holes.

References

[1] C. Choffrut, J. Karhumäki, Combinatorics on words, in: G. Rozenberg, A. Salomaa (Eds.), Handbook of Formal Languages, vol. 1, Ch. 6, Springer, Berlin, 1997.

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