INFORMATIQUE THÉORIQUE ET APPLICATIONS

ALEXANDRU MATEESCU
ARTO SALOMAA

PCP-prime words and primality types

Informatique théorique et applications, tome 27, nº 1 (1993), p. 57-70.
<http://www.numdam.org/item?id=ITA_1993__27_1_57_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 27, n° 1, 1993, p. 57 à 70)

PCP-PRIME WORDS AND PRIMALITY TYPES (*)

by Alexandru MATEESCU [1] and Arto SALOMAA [2]

Communicated by J. BERSTEL

Abstract. – *We investigate "simplest", "primitive" or "prime" solutions of instances of the Post Correspondence Problem, PCP. We take also the opposite view point by studying words that can describe such a prime solution, for some instance of PCP.*

Résumé. – *Nous étudions les solutions « les plus simples », « primitives » ou « premières » d'instances du problème de correspondance de Post (PCP). Nous adoptons également le point de vue opposé en étudiant les mots qui peuvent décrire une telle solution première.*

1. INTRODUCTION AND PREVIOUS RESULTS

Post Correspondence Problem, [3], is one of the very basic undecidable problems. Reduction to PCP is the most common one in language theory. Whenever w_1 and w_2 are solutions for an instance of PCP, then so is $w_1 w_2$. It is natural to consider w_1 and w_2 to be "simpler" or "more primitive" solutions than $w_1 w_2$. Such considerations have also turned out to be theoretically important in various contexts, see [4].

This paper continues the systematic study of "primitive" or "prime" solutions initiated in [5]. We first review the basic definitions.

(*) Received July 1991, accepted December 1991.

[1] Current address: University of Bucharest, Faculty of Mathematics, Academiei 14, sector 1, R-70109 Bucharest, Romania. University of Joensuu, Department of Computer Science, SF-80101 Joensuu, Finland.

[2] All correspondence to Arto Salomaa: Academy of Finland and Mathematics Department, University of Turku, SF-20500 Turku, Finland.

Informatique théorique et Applications/Theoretical Informatics and Applications
0988-3754/93/01 57 14/$3.40/© AFCET-Gauthier-Villars

Let g and h be nonerasing morphisms of Σ^* into Δ^*, where Σ and Δ are finite alphabets. The *equality set* between g and h is defined by

$$E(g, h)=\{ w\in\Sigma^+ \mid g(w)=h(w) \}.$$

(Observe that the empty word λ is not considered to be a member of the equality set.) The pair $(g, h)=\mathrm{PCP}$ is also refered to as an instance of the Post Correspondence Problem. Words in $E(g, h)$, if any are called *solutions* of PCP.

For a word w over Σ, we now consider the sets of words obtained from w by removing a final subword, a subword, or a scattered subword, respectively. By definition,

$$\mathrm{fin}(w)=\{ v_1 \mid w=v_1 v_2, \text{ for some } v_2\in\Sigma^* \},$$

$$\mathrm{sub}(w)=\{ v_1 v_2 \mid w=v_1 x v_2, \text{ for some } v_1, v_2, x\in\Sigma^* \},$$

$$\mathrm{scatsub}(w)=\{ v_1 \ldots v_k \mid w=x_1 v_1 \ldots x_k v_k x_{k+1}, \text{ for some } x_i, v_i\in\Sigma^* \}.$$

Three further sets determined by the pair (g, h) are now defined as follows:

$$F(g, h)=\{ w\in E(g, h) \mid \mathrm{fin}(w)\cap E(g, h)=\{ w \} \},$$

$$S(g, h)=\{ w\in E(g, h) \mid \mathrm{sub}(w)\cap E(g, h)=\{ w \} \},$$

$$P(g, h)=\{ w\in E(g, h) \mid \mathrm{scatsub}(w)\cap E(g, h)=\{ w \} \}.$$

Words in the three sets are called *F-prime*, *S-prime* and *prime solutions* for the instance $\mathrm{PCP}=(g, h)$, respectively.

It is a direct consequence of the definitions that

$$P(g, h)\subseteq S(g, h)\subseteq F(g, h)\subseteq E(g, h).$$

One or both of the first two inclusions may be strict, whereas the third inclusion is always strict, provided $E(g, h)$ is nonempty. If $E(g, h)$ is nonempty then so are the three other sets. Clearly, each of the four sets is recursive.

The triple (p, s, f), where p, s, and f are the cardinalities of the sets $P(g, h)$, $S(g, h)$, and $F(g, h)$, respectively, is defined to be the *primality type* of the instance $\mathrm{PCP}=(g, h)$. Thus, p, s, and f are non-negative integers or ∞.

The following three results were established in [5].

THEOREM 1: *If $E(g, h)$ is a regular language, then $S(g, h)$ is finite.*

LEMMA 1: *Assume that the word xyz is in $F(g, h)$, where x, y, z are nonempty, and that xz is in $E(g, h)$. Then $xy^n z$ is in $F(g, h)$, for all $n\geqq 0$.*

THEOREM 2: *A triple (p, s, f) is a primality type iff either $p=s=f=0$, or else (i) $1 \leqq p \leqq s \leqq f$, (ii) p is finite, and (iii) if $s<f$ then $f=\infty$. An example for each possible type can be effectively constructed.*

Also the following lemma, originally due to [2] and basic in the combinatorics of words, will be needed in sequel.

LEMMA 2: *If $uv=vz$ holds for some words u, v, z, where u is nonempty, then*

$$u=xy, \qquad v=(xy)^k x, \qquad z=yx,$$

for some words x, y and integer $k \geqq 0$. If $uv=vu$ holds for some nonempty words u and v, then u and v are powers of the some word.

2. A MORE GENERAL SETUP

For many purposes it is useful to view the languages $P(g, h)$, $S(g, h)$ and $F(g, h)$ introduced above as subsets of the equality set $E(g, h)$, obtained from $E(g, h)$ by certain operations.

Consider a partial order $\leqq$ on Σ^* and a language $L \subseteqq \Sigma^*$. Then $\mathrm{MIN}_{\leqq}(L)$ is the subset of L consisting of elements minimal with respect to $\leqq$, in symbols,

$$\mathrm{MIN}_{\leqq}(L)=\{x \in L \mid \text{whenever } y \in L \text{ satisfies } y \leqq x, \text{ then } y=x\}.$$

Consider, further, the following binary relations in Σ^*:

$$y F w \text{ iff } y \text{ is in fin}(w),$$

$$y S w \text{ iff } y \text{ is in sub}(w),$$

$$y P w \text{ iff } y \text{ is in scatsub}(w),$$

$$y \mathrm{D} w \text{ iff } w=xyz, \text{ for some } x \text{ and } z.$$

Observe that the notations F, S and P used in the relations are in accordance with the notations $F(g, h)$, $S(g, h)$ and $P(g, h)$. The relations S and D are dual in the sense that D is the usual subword relation whereas S indicates what is left over when a subword is removed. The relation P is self-dual in the same sense and, consequently,

$$P(g, h)=\mathrm{MIN}_P(E(g, h)).$$

Clearly, F, P and D are partial orders and, thus, we have also

$$F(g, h)=\mathrm{MIN}_F(E(g, h)).$$

As regards S, the situation is trickier. The relation S is reflexive and antisymmetric but not transitive and, consequently, not a partial order. For instance, we have *acSabc* and *abcSabcb* but not *acSabcb*. However, the following theorem holds – the proof is left to the reader.

THEOREM 3: *The transitive closure S^+ of S is a partial order and, moreover, $S^+=P$.*

The inclusions between the different families are depicted by the following diagram.

$$\begin{tikzcd}
& E(g, h) & \\
& F(g, h)=\mathrm{MIN}_F(E(g, h)) \arrow[u] & \\
S(g, h) \arrow[ur] & & \mathrm{MIN}_D(E(g, h)) \arrow[ul] \\
& P(g, h)=\mathrm{MIN}_P(E(g, h)) \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Each of the inclusions may be strict. That the sets $S(g, h)$ and $\mathrm{MIN}_D(E(g, h))$ are incomparable is shown by the following examples. We use numerals as letters of Σ to point out the customary definition of the Post Correspondence Problem as two lists of words. The first example is also historically interesting, a slight modification of the example given in [3].

Let g and h be defined by the table

	1	2	3
g	bb	ab	e
h	b	ba	be

Now $S(g, h)=\{13\}$ but $\mathrm{MIN}_D(E(g, h))=12^*3$.

Secondly, define g and h by

	1	2	3	4
g	$(ab)^3$	a	$(ba)^2$	b
h	ab	aba	ba	$b(ab)^2$

Now 1234 is in $S(g, h)$ but not in $\mathrm{MIN}_D(E(g, h))$, since 23 is in $E(g, h)$.

Many closure properties can be obtained for the operations MIN_P, MIN_D and MIN_F in a fairly straightforward manner. Some of the properties are useful in the study of *P*-, *S*-and *F-languages.* The following theorem does not intend to give an exhaustive list.

THEOREM 4: *The families of regular, context-sensitive and recursive languages are all closed under the operation* MIN_F. *The families of context-free and recursively enumerable languages are not closed under* MIN_F. *In general, an AFL is closed under* MIN_F *iff it is closed under complementation.*

Proof: By known closure properties (also the recently established closure of the family of context-sensitive languages under complementation), it suffices to prove the last sentence. Because of the equation

$$\mathrm{MIN}_F(L)=L\cap\sim(L\Sigma^+),$$

an AFL closed under complementation is closed under MIN_F. (Clearly, closure under complementation implies closure under intersection is well.) Conversely, let $\mathscr{L}$ be an AFL closed under MIN_F and let $L\subseteqq\Sigma^*$ be in $\mathscr{L}$. Moreover, choose two letters c and d not in Σ, and consider the morphism h mapping letters of Σ into themselves and c and d to the empty word λ. Because $\sim L=h(L_1)$, where

$$L_1=(\mathrm{MIN}_F(Lc\cup\Sigma^*cd))\cap\Sigma^*cd,$$

and h erases only two symbols from each word, we conclude that $\mathscr{L}$ is closed under complementation. Indeed, L_1 is obtained by adding the suffix cd to words belonging to the complement of L. □

Comment: Let's consider the languages: $L_1=\{a^ib^jc^j\mid i,j\geqq 1\}$, $L_2=\{a^kb^k\mid k\geqq 1\}$ and $L=L_1\cup L_2$.

Remark that $a^+b^+c^+\cap\mathrm{MIN}_F(L)=\{a^nb^mc^m\mid n>m\geqq 1\}$ and hence $\mathrm{MIN}_F(L)$ is not a context-free language, despite that L_1 and L_2 are deterministic context-free languages.

In addition to theorem 1, many other results can be obtained concerning the interrelations between the various languages associated with the PCP-instance (g, h). The numbers s and f refer to the components of the primamity type in the next theorem.

THEOREM 5: $E(g, h)=(F(g, h))^+$ *and, moreover, every word of* $E(g, h)$ *has a unique decomposition in terms of words in* $F(g, h)$. *The language* $E(g, h)$ *is regular iff* $F(g, h)$ *is regular. If* f *is finite then* $E(g, h)$ *is regular. The language* $S(g, h)$ *is finite iff, for some* k, *every word* w *in* $E(g, h)$ *with length greater* *than k can be represented as $w=xyz$, where y is not empty and $xy^i z$ is in $E(g, h)$, for all i.*

The rather straightforward proof of Theorem 5 is omitted. Lemma 1 (which, in fact, gives a stronger result) should be used in establishing the last sentence of Theorem 5.

It is an open problem what can be concluded about $E(g, h)$ if s is finite. The finiteness of s does not imply that $E(g, h)$ is regular. This is seen from the morphisms $g(a)=h(b)=a$, $g(b)=h(a)=a^2$. Does it imply, for instance, that $E(g, h)$ is context-free?

3. AN EXTENSION AND ITS INFLUENCE ON THE PRIMALITY TYPE

Consider an instance (g, h) of the PCP and a natural number $q \geqq 2$. We assume that the instance is *nontrivial* in the sense that $g(a)=h(a)$ holds for no letter a of the alphabet Σ. We denote

$$\Sigma_q = \Sigma^2 \cup \Sigma^3 \cup \ldots \cup \Sigma^{q+1},$$

where the right side is understood as an alphabet whose letters are denoted by

$$[b_1 b_2 \ldots b_t], \quad 2 \leqq t \leqq q+1, \quad b_i \in \Sigma.$$

Finally, define two morphisms $g', h' : \Sigma^*_q \to \Delta^*$, by

$$g'([b_1 b_2 \ldots b_t]) = g(b_1) g(b_2) \ldots g(b_t),$$
$$h'([b_1 b_2 \ldots b_t]) = h(b_1) h(b_2) \ldots h(b_t).$$

Then the pair (g', h') is also an instance of the PCP, referred to as the *q-extension* of the instance (g, h).

The following lemma is an immediate consequence of the definitions.

LEMMA 3: *Let (g', h') and (g'', h'') be q- and r-extensions, respectively, of (g, h) with $q<r$. Then*

$$P(g', h') \subseteq P(g'', h''), \qquad S(g', h') \subseteq S(g'', h'') \qquad \text{and} \qquad F(g', h') \subseteq F(g'', h'').$$

We note in passing that it is decidable whether or not an instance of the PCP is a q-extension of some other instance. The next theorem shows that the component f in the primality type is always 0 or ∞ if attention is restricted to q-extensions.

THEOREM 6: *If $E(g, h)$ is nonempty then $F(g', h')$ is infinite for every q-extensions (g', h') of (g, h).*

Proof: By Lemma 3, it suffices to consider the case $q=2$. Let $i_1 i_2 \ldots i_k$ be an F-prime solution of the instance (g, h). Because we assume in this section that the instances are nontrivial, we have $k \geqq 2$. Starting from the given F-prime solution, we construct an infinite set of F-prime solutions of the q-extension.

If $k>3$, such a set is

$$[i_1 i_2][i_3 i_4] \ldots [i_{k-2} i_{k-1}]([i_k i_1][i_2 i_3] \ldots [i_{k-5} i_{k-4}][i_{k-3} i_{k-2} i_{k-1}])^* \\ [i_k i_1][i_2 i_3] \ldots [i_{k-1} i_k]$$

if k is odd, and

$$[i_1 i_2][i_3 i_4] \ldots [i_{k-5} i_{k-4}][i_{k-3} i_{k-2} i_{k-1}]([i_k i_1][i_2 i_3] \ldots [i_{k-2} i_{k-1}])^* \\ [i_k i_1][i_2 i_3][\ldots [i_{k-2} i_{k-1} i_k]$$

if k is even.

If $k=3$ or $k=2$, such a set is

$$[i_1 i_2]([i_3 i_1 i_2])^* [i_3 i_1][i_2 i_3]$$

or

$$[i_1 i_2 i_1][([i_2 i_1])^* [i_2 i_1 i_2]. \quad \square$$

Consider the following example. For the instance (g, h) defined by

	1	2
g	ab	a
h	a	ba

we have $P(g, h)=S(g, h)=F(g, h)=\{12\}$ and, hence, the primality type of (g, h) is $(1, 1, 1)$. For the 2-extension (g', h'), we have

$$[121]([21])^*[212] \subseteqq F(g', h')$$

and

$$S(g', h')=P(g', h')=\{[12], [121][212]\}.$$

The primality type of (g', h') is $(2, 2, \infty)$.

The equality sets $E(g, h)$ and $E(g', h')$ where (g', h') is the q-extension of (g, h) are *gsm*-equivalent in the sense of the following theorem. The standard *gsm*-construction, where states are provided with buffers, is left to the reader. Observe that M is in general nondeterministic and erasing.

THEOREM 7: *Let* (g, h) *be an arbitrary instance of the PCP and* (g', h') *the* q*-extension of* (g, h), *for some* $q \geqq 2$. *Then there are generalized sequential machines* M *and* M' *with the property*

$$E(g, h)=M'(E(g', h')) \text{ and } E(g', h')=M(E(g, h)).$$

The following result is an immediate consequence of Theorems 1 and 7.

THEOREM 8: *Assume that* $E(g, h)$ *is regular. Then* $S(g', h')$ *is finite, for every* q*-extension* (g', h') *of* (g, h).

In fact, Theorem 8 can be expressed more generally as follows. Assume that a certain property P of $E(g, h)$ implies a certain other property Q of $S(g, h)$ and that, moreover, the former property P is preserved under *gsm*-mappings. Then if $E(g, h)$ possesses P, $S(g', h')$ possesses Q, where (g', h') is a q-extension of (g, h).

According to theorem 6, the transition to a q-extension always makes the component in the primality type infinite (provided $f>0$ originally). According to Theorem 8, there are instances such that no such transition makes the component s infinite. However, in some cases the transition to the q-extension makes s indeed infinite, as seen below.

For the instance (g, h) defined by

	1	2
g	a^2	a
h	a	a^2

$F(g, h)$ is infinite but $S(g, h)=\{12, 21\}$. Since words of the form $[111]\,[11]^i\,[122]\,[22]^{i+1}$ are in $S(g', h')$ we conclude that $S(g', h')$ is infinite, for every q-extension (g', h').

4. PRIME WORDS AND LANGUAGES

We now take the opposite point of view. We consider arbitrary words and languages and ask whether they can appear as some type of prime solutions, for some instance (g, h) of the PCP. For example, the word *abab* cannot be a prime solution (of any type) for any instance (g, h).

We say that a word w over Σ is *P-prime* if, for some (g, h), w is in $P(g, h)$. *S-prime* and *F-prime* words are defined similarly. In the sequel we speak, briefly, of *P-words*, *S-words* and *F-words*.

Similarly, we may define *P-languages*, *S-languages* and *F-languages*. In fact, these notions can be defined in two different ways, depending on whether we consider inclusion or equality. Thus, L is a P-language in the *first* (respective *second*) *sense* if, for some instance (g, h), L is included in (respective equals) $P(g, h)$. Clearly, if *abab* is a word in L, then L cannot be a P-language in either sense (and also not an S- or F-language). We restrict our attention here to words and hope to return to languages in another paper.

Consider, thus, words over the alphabet Σ. If Σ consists only of one letter a, then only the word a is prime, no matter which of the three types we consider. Therefore, we assume in the sequel that

$$\Sigma=\{a_1, \ldots, a_n\}, \qquad n \geqq 2.$$

We make the convention that when we speak of a word w over Σ then all letters of Σ actually occur in w. The *Parikh vector* associated to w is denoted by $\psi(w)$.

Clearly, the set of P-words is included in the set of S-words which, in turn, is included in the set of F-words. The following lemma gives a way of constructing words that are not F-words.

LEMMA 4: *If w_1 is a nontrivial prefix of w satisfying $\psi(w_1)=r\psi(w)$, for some (rational) number r, then w is not an F-word.*

Proof: Observe first that, for any word x and pair (g, h), the "balance"

$$|g(x)|-|h(x)|$$

of x with respect to (g, h) depends only in $\psi(x)$. (In other words, x_1 and x_2 have the same balance if $\psi(x_1)=\psi(x_2)$.)

If w were an F-word with (g, h) being the instance in question, then w would be in $E(g, h)$ and consequently, the balance of w would be 0. But then also the balance of w_1 would be 0 and w_1 would be in $E(g, h)$, a contradiction. □

If w is not an F-word, it cannot be an S- or P-word either. Lemma 4 gives, among others, the following examples:

$$w^i,\ i>1; \qquad w=w_1 \ldots w_k,\quad k>1,\quad \text{if } \psi(w_1)=\psi(w_2)=\ldots=\psi(w_k);$$
$$ab^2a^2b^4;\ ab^6a^3b^2;\ abcab^5a^2c.$$

Following the terminology of [1], we say that a language L is *rich* if there is no nontrivial instance (g, h) such that $L \subseteq E(g, h)$.

(Recall that the instance being trivial means that $g(a)=h(a)$ holds for some letter a.)

Clearly, no singleton $L=\{w\}$ is rich because we can easily construct a nontrivial instance (g, h) with only one letter in the target alphabet Δ such that $g(w)=h(w)$.

A morphism h is *periodic* if all values $h(w)$ are powers of a single word, that is, there is a word u, referred to as the *period* of h, such that for each letter a there is an integer i with the property $h(a)=u^i$. The period is unique if we consider the word with the minimal length.

A language L is *almost rich* if, whenever $L \subseteq E(g, h)$ holds for a nontrivial instance (g, h), then both g and h are periodic. A word w is *almost rich* if $L=\{w\}$ is almost rich. The following lemma established in [5] shows that almost richness implies that g and h are periodic with the *same* period.

LEMMA 5: *Assume that g and h are periodic and $E(g, h)$ is not empty. Then g and h are periodic with the same period.*

LEMMA 6: *Assume that $w=a^2bvab^2$ where v is a word over $\{a, b\}$ containing equally many a's and b's in such a way that no prefix of v contains more b's than a's. Then w is an F-word but not an S-word.*

Proof: The instance (g, h) defined at the end of section 3 shows that w is an F-word. The assumptions guarantee that the balance 0 is reached only at the end. (In fact, the assumptions concerning v mean that v is in the Dyck language.)

Assume now that w is an S-word and that (g, h) is the instance showing it. Clearly, (g, h) must be nontrivial. We know that w is in $E(g, h)$. Considering the beginning and the end of w, we see that we can assume without loss of generality that g and h are defined by

	a	b
g	t	yu
h	tx	u

where t, u, x and y are nonempty and $|x|=|y|$.

Since w is in $E(g, h)$, we obtain

$$ttyug(v)\,tyuyu=txtxuh(v)\,txuu$$

Reading the words from the left and from the right, we obtain

$$ty=xt \qquad \text{and} \qquad uy=xu.$$

Consequently,

$$g(aabb)=ttuyu=txtuyu=txtxuu=h(aabb).$$

Thus, $a^2 b^2$ is in $E(g, h)$. This contradiction shows that w is not an S-word. □

LEMMA 7: *Words of the form $w=avava$, where a is a letter and v is over the alphabet $\{a, b\}$, are almost rich.*

Proof: We assume that $g(w)=h(w)$, where (g, h) is nontrivial. We cannot have $g(av)=h(av)$ because we would then have also $g(a)=h(a)$, by our assumption $g(w)=h(w)$. Without loss of generality, we assume that

$$g(av)=\alpha, \qquad h(av)=\beta=\alpha x, \qquad |x|\geqq 1.$$

We denote also

$$g(a)=\alpha_1,\ h(a)=\beta_1, \qquad \alpha=\alpha_1\alpha_2, \qquad \beta=\beta_1\beta_2.$$

It follows that $|x|<|\beta_2|$ and, moreover,

$$\alpha_1=\beta_1 y, \qquad |y|\geqq 1.$$

We now write the equation $g(w)=h(w)$ in the new notation:

$$\beta_1 y\alpha_2\beta_1 y\alpha_2\beta_1 y=\beta_1\beta_2\beta_1\beta_2\beta_1=\beta_1 y\alpha_2 x\beta_1 y\alpha_2 x\beta_1.$$

Consequently,

$$\beta_1 y\alpha_2\beta_1 y=x\beta_1 y\alpha_2 x\beta_1.$$

This implies that $|y|=2.|x|$. We write

$$y=y_1 y_2 \qquad \text{with} \quad |y_1|=|y_2|=|x|.$$

Our preceding equation reads now

$$\beta_1 y_1 y_2\alpha_2\beta_1 y_1 y_2=x\beta_1 y_1 y_2\alpha_2 x\beta_1.$$

Considering subwords of the same length, we obtain

$$\beta_1 y_1=x\beta_1, \qquad y_1=y_2, \qquad \alpha_2\beta_1 y_1 y_1=y_1\alpha_2 x\beta_1,$$

where we have already used the second equation in the third equation.

We now apply Lemma 2. The first equation above yields the representation

$$\beta_1=(uv_1)^r, \qquad x=uv_1, \qquad y_1=v_1 u.$$

Using this, we obtain from the last equation above

$$\alpha_2 uv_1 = v_1 u \alpha_2,$$

which yields another representation

$$\alpha_2 = (rs)^1 r, \qquad v_1 u = rs, \qquad uv_1 = sr.$$

Writing, without loss of generality, $v_1 = rr_1$, we obtain $s = r_1 u$ and $urr_1 = r_1 ur$. By Lemma 2,

$$r_1 = z^i, \qquad ur = z^j,$$

for some z. Substituing everything in the equation $g(w) = h(w)$ and dividing from left and right, we are left with the equation $uv_1 z^j = z^j uv_1$ which shows that uv_1 and z^j are powers of the same word. Combining our knowledge, we see that all of α_1, α_2, β_1, β_2 are powers of the same word, hence the lemma follows. □

It is likely that the lemma holds also if v is over an arbitrary alphabet. The lemmas gives a possibilitly of constructing (by using the three marked occurrences of a as boundaries) S-words that are not P-words. However, we have no explicit examples, and such a construction remains an open problem.

We now prove some results concerning P-words. We remind the reader of the conventions concerning Σ, made at the begin of this section.

THEOREM 9: *For every n-dimensional vector v with positive integer components, there is a P-word w such that $\psi(w) = v$.*

Proof: Denote $v = (i_1, i_2, \ldots, i_n)$. We'll prove that

$$w = a_1^{i_1} a_2^{i_2} \ldots a_n^{i_n}$$

is a P-word.

Let k be the least common multiple of the numbers $i_1, i_2, \ldots, i_n$. Denote

$$j_t = k/i_t, \qquad t = 1, \ldots, n, \qquad \text{and} \qquad r = i_1 i_n.$$

Consider the instance (g, h) defined by

	a_1	a_2	$\ldots$	a_{n-1}	a_n
g	$(c^r d_1)^{j_1}$	$(c^r d_2)^{j_2}$	$\ldots$	$(c^r d_{n-1})^{j_{n-1}}$	c^{i_1}
h	c^{i_n}	$(d_1 c^r)^{j_2}$	$\ldots$	$(d_{n-2} c^r)^{j_{n-1}}$	$(d_{n-1} c^r)^{j_n}$

where $\Delta = \{ c, d_1, \ldots, d_{n-1} \}$. We compute:

$$\begin{aligned} g(w) &= (c^r d_1)^{j_1 i_1} (c^r d_2)^{j_2 i_2} \ldots (c^r d_{n-1})^{j_{n-1} i_{n-1}} c^{i_1 i_n} \\ &= (c^r d_1)^k (c^r d_2)^k \ldots (c^r d_{n-1})^k c^r \\ &= c^r (d_1 c^r)^k (d_2 c^r)^k \ldots (d_{n-1} c^r)^k \\ &= c^{i_n i_1} (d_1 c^r)^{j_2 i_2} \ldots (d_{n-1} c^r)^{j_n i_n} = h(a_1^{i_1} a_2^{i_2} \ldots a_n^{i_n}) = h(w). \end{aligned}$$

It is not difficult to see that if something is removed from w, the resulting word is not any more in the equality set. Therefore, w is a P-word. □

There are many other P-words than those given in Theorem 9. We have the following general result.

THEOREM 10: *Every word w over the alphabet $\{a_1, a_2\}$, such that the two components i_1 and i_2 in $\psi(w)$ are relatively prime, is a P-word. More generally, let w be over $\Sigma = \{a_1, \ldots, a_n\}$ and assume that Σ can be divided into two disjoint nonempty parts Σ_1 and Σ_2 such that the two numbers, obtained by summing up the components of $\psi(w)$ corresponding to the letters of Σ_1, as well as those corresponding to the letters of Σ_2, are relatively prime.*

Proof: Consider the first sentence. Define the instance (g, h) by

	a_1	a_2
g	b	b^{i_1+1}
h	b^{i_2+1}	b

Read the word w from left to right and look at the balance. After reading the prefix w_1, the balance

$$s i_1 - r i_2$$

is observed, if $\psi(w_1) = (r, s)$. The equation

$$s i_1 - r i_2 = 0,$$

where $r < i_1$ or $s < i_2$, would contradict i_1 and i_2 being relatively prime. Thus, no suffix can be removed from w and still stay in the equality set. The same argument concerns the removal of scattered subwords because Δ contains only one letter. This proves the first sentence.

The second sentence is an immediate consequence: just identify the letters of Σ_1, as well those of Σ_2, in the definition of the instance (g, h). □

Observe that the components of the Parikh vector can be pairwise mutually prime, and the condition of the second sentence is still not satisfied. A counter example is (3, 5, 22).

Our final lemma is a useful tool for constructing words that are not P-words.

LEMMA 8: *If w is not a P-word and φ is a nonerasing morphism, then $\varphi(w)$ is not a P-word.*

Proof: Assume the contrary: for some instance (g, h), φ (w) is in $P(g, h)$. Consequently, w is in $E(g\varphi, h\varphi)$. Since w is not in $P(g\varphi, h\varphi)$, there is a word y in $E(g\varphi, h\varphi)$ satisfying $y P w$, where P is the relation defined in Section 2. This implies that $\varphi(y) \mathrm{P} \varphi(w)$ and $\varphi(y) \in E(g, h)$. (Since φ is nonerasing, $\varphi(y)$ is not the empty word.) Thus, $\varphi(w)$ is not in $P(g, h)$, a contradiction. □

By Lemmas 6 and 8 we may conclude, for example, that $a^4 b^3 a^2 b^6$ is not a P-word. However, we want to emphasize that our results do not give an exhaustive characterization of P-words.

5. CONCLUSION

We have investigated some issues fundamental for the Post Correspondence Problem and equality sets and, more generally, for the basic combinatorics of words. Many of the problems remain open. We have not discussed at all the area of prime languages (as defined in Section 4). It is also a challenging problem to estimate the component p in the primality type, for example, in terms of the cardinality of Σ.

ACKNOWLEDGEMENTS

We thank the referee for pointing out errors in the original version of Theorem 4.

REFERENCES

1. K. CULIK II and A. SALOMAA, Test Sets and Checking Words for Homomorphism Equivalence, *J. Comput. System Sci.*, 1980, *20*, pp. 379-395.
2. R. C. LYNDON and M. P. SCHUTZENBERGER, The Equation $a^M = b^N c^P$ in Free Group, *Michigan Math. J.*, 1962, *9*, pp. 289-298.
3. E. POST, A Variant of a Recursively Unsolvable Problem, *Bull. Amer. Math. Soc.*, 1946, *52*, pp. 264-268.
4. A. SALOMAA, Jewels of Formal Language Theory, *Comput. Sci. Press*, 1981.
5. A. SALOMAA, K. SALOMAA and SHENG YU, Primality types of instances of the Post Correspondence Problem, *E.A.T.C.S. Bulletin*, 1991, *44*.