# Decision Problems Concerning Prime Words and Languages of the PCP

**Marjo Lipponen**

Turku Centre for Computer Science
TUCS Technical Report No 27
June 1996
ISBN 951-650-783-2
ISSN 1239-1891

**Abstract**

This paper investigates properties of prime words and prime languages obtained from the Post Correspondence Problem. We show that the properties of being a prime word or a finite prime language are decidable. We also present other characterization methods.



**TUCS Research Group**
Mathematical Structures of Computer Science



[1]To be presented at the 8th International Conference on Automata and Formal Languages, Salgótarján, Hungary, July 29 – August 2, 1996

# 1 Prime solutions of the PCP

It has become customary, starting from [14], to consider three types of solutions for an instance of the Post Correspondence Problem that are somehow simpler than the other solutions. A solution is termed *F-prime* if no (nonempty) final subword can be removed such that what remains is still a solution. In the same way *S-prime* and *P-prime solutions* correspond to removing a subword and a scattered subword, respectively.

More specifically, by an *instance* of the Post Correspondence Problem we mean a pair $(g, h)$ of nonerasing morphisms $g, h : \Sigma^* \to \Delta^*$ where $\Sigma$ and $\Delta$ are finite alphabets, and by its *solutions* the words in the *equality set*

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

Hence a solution is a nonempty word whose images under $g$ and $h$ coincide. If $g(a) = h(a)$ for some $a \in \Sigma$, we call such an instance $(g, h)$ *trivial* and if there exists $u \in \Delta^+$ such that $g(a) = u^{i_a}$ and $h(a) = u^{j_a}$ for all $a \in \Sigma$ then $(g, h)$ is called *periodic*.

The sets of F-prime, S-prime and P-prime solutions, based on removing a final subword, a subword or a scattered subword, are defined by

$$\begin{aligned} F(g,h) &= \{w \in E(g,h) \mid \mathit{fin}(w) \cap E(g,h) = \{w\}\}, \\ S(g,h) &= \{w \in E(g,h) \mid \mathit{sub}(w) \cap E(g,h) = \{w\}\}, \\ P(g,h) &= \{w \in E(g,h) \mid \mathit{scatsub}(w) \cap E(g,h) = \{w\}\}, \end{aligned}$$

where

$$\begin{aligned} \mathit{fin}(w) &= \{v \mid w = vx, \text{ for some } x \in \Sigma^*\}, \\ \mathit{sub}(w) &= \{v_1 v_2 \mid w = v_1 x v_2, \text{ for some } v_1, v_2, x \in \Sigma^*\}, \\ \mathit{scatsub}(w) &= \{v_1 \dots v_k \mid w = x_1 v_1 \dots x_k v_k x_{k+1}, \text{ for some } x_i, v_i \in \Sigma^*\}. \end{aligned}$$

The study of prime words, initiated in [9] and continued in [3–7], is an extension of that of prime solutions: every word that is F-|S-|P-prime for some instance $(g, h)$ of the Post Correspondence Problem is an *F-|S-|P-word*. Since every P-prime is an S-prime and every S-prime is an F-prime, also prime words form an increasing hierarchy. The details of this study can be found in [4, 5].

Prime languages can be defined in two ways, based on equality or inclusion. We say that a language $L$ is an *F-, S-* or a *P-language* if, for some instance $(g, h)$, $L$ either equals or is a part of the set $P(g, h)$, $S(g, h)$ or $F(g, h)$, respectively.

Though the Post Correspondence Problem, [10], is most famous for its use in undecidability proofs, [11], many properties of prime words and languages turn out to be decidable.

We study the membership problems of prime words and languages as well as other methods to characterize them. The last section is devoted to the binary alphabet.

The results we are presenting in this paper are based on a dissertation, [6], published only recently.

For further details of formal language theory we refer to [12].

## 2 Decision problems

We start by showing that for any word $w$ we can decide whether $w$ is a P-prime, an S-prime or an F-prime solution for some instance $(g, h)$.

**Theorem 2.1** *Each of the properties of being a P-word, an S-word or an F-word is decidable.*

**Proof:** To check whether a given word $w$ is a P-word we have to consider the set $scatsub(w)$, its scattered subwords, which is always finite. Using the result of Makanin [8] and the possibility to translate an inequality into several systems of equations as in [1] and [2], we can test for each of the following systems whether it has a solution, i.e., whether there are $g$ and $h$ such that

$$\begin{aligned} &g(w) = h(w) \\ &g(v) \neq h(v) \quad \text{for all } v \in scatsub(w) - \{w, \lambda\}. \end{aligned}$$

If so, then $w$ is a P-prime solution for this instance $(g, h)$, otherwise $w$ is not a P-word.

With S-words and F-words we use similarly the sets $sub(w)$ and $fin(w)$, respectively. □

This first result tells actually very little of the nature of the three types of prime words since Makanin's general algorithm is very complicated. This is why we seek for other possibilities to characterize these words.

We start with two important notions. A *basic Parikh vector* $\psi_0$ is obtained from the Parikh vector $\psi$ by dividing it with the greatest common divisor of its components. Parikh vectors can also be used to make comparisons between words. We say that a word $u$ is *Parikh shorter* than $v$ if $\psi(u) \leq \psi(v)$ componentwise.

The following result was established in [5]. It gives us an effective tool for characterizing prime words.

**Lemma 2.2** *For each word $w$ we can effectively find an instance of the Post Correspondence Problem such that for all words $w'$ which are Parikh shorter than $w$, $w'$ is also a solution if and only if $\psi_0(w') = \psi_0(w)$.*

A word is said to be *ratioprimitive* (resp. *subratioprimitive*) if none of its proper prefixes (resp. subwords) has the same basic Parikh vector as the whole word.

**Theorem 2.3** *A nonempty word $w$ is an F-word if and only if it is ratioprimitive.*

Theorem 2.3 gives an effective algorithm for F-words: it is easy to check for a given word whether its prefixes have the same basic Parikh vector as the whole word, even in polynomial time.

For S-words we have found only a partial algorithm.

**Theorem 2.4** *Every subratioprimitive word is an S-word.*

The converse of the previous theorem does not hold; for instance, the word 311132223 is an S-word but not subratioprimitive.

Also for P-words the algorithm is partial. The following theorem shows what kind of words can appear as P-words. The first result follows directly from Lemma 2.2 and the second one is due to [9].

**Theorem 2.5** *The word $w$ is a P-word if*

1. *$\psi(w) = \psi_0(w)$ or*
2. *$w = a_1^{i_1} a_2^{i_2} \dots a_n^{i_n}$ where $\Sigma = \{a_1, \dots, a_n\}$ and $n \geq 2$.*

These results are not exhaustive either, for instance, the words 121233, 122313, 122133, 122331, 121323 are all P-words.

These results can be improved, however, with a new restriction. We say that a word $w$ is *periodicity forcing* if every instance $(g, h)$ for which $w$ is in $E(g, h)$ is periodic or trivial.

**Lemma 2.6** *If $w$ is periodicity forcing and $\psi(w) \neq \psi_0(w)$ (resp. not subratioprimitive) then $w$ is not a P-word (resp. an S-word).*

Unfortunately, periodicity forcing words are not characterized any more than prime words, see [6] for details. On the other hand, Lemma 2.6 is not true for the other direction; for instance, the word 1212123123 which is not subratioprimitive is not an S-word either (see [4]) even though it is not periodicity forcing.

The following result of prime languages can be viewed as an extension of Theorem 2.1.

**Theorem 2.7** *For finite languages each of the properties of being a P-, an S- and an F-language in inclusion sense is decidable.*

**Proof:** In order to decide whether a given language $L = \{w_1, \ldots, w_n\}$ is a P-language we apply Makanin's result, [8], for the finite system of equations,

$$\begin{aligned}
g(w_1) &= h(w_1), \\
&\vdots \\
g(w_n) &= h(w_n), \\
g(u_1) &\neq h(u_1) \quad \text{for all } u_1 \in scatsub(w_1) - \{w_1, \lambda\}, \\
&\vdots \\
g(u_n) &\neq h(u_n) \quad \text{for all } u_n \in scatsub(w_n) - \{w_n, \lambda\}.
\end{aligned} \tag{1}$$

If this system has a solution; that is, the equations hold for some $g$ and $h$, then $L$ is a subset of $P(g,h)$; otherwise, $L$ cannot be a P-language.

For S- or F-languages we use similarly the sets *sub* and *fin*. □

What is the situation with infinite languages? Since P-languages are always finite (see [14]), they are decidable also in this case. On the other hand, with S- and F-languages we cannot apply the previous argument any more, the system of equations being infinite.

By Ehrenfeucht's conjecture (see [13] for details) every language $L$ possesses a finite subset $D$, called a test set, such that, whenever $g$ and $h$ are two morphisms defined on $\Sigma^*$ and satisfying $g(w) = h(w)$ for every $w$ in $D$, then $g(w) = h(w)$ holds for every $w$ in $L$. The construction of $D$ is not effective in general but is effective, for instance, for context-free languages. With this in mind we cannot, however, be sure that all the members of $L$ are F- or S-prime solutions even if this is the case for $D$.

In the same context, [13], a more general result was proved: Every system of word equations possesses a finite subsystem equivalent to the original system. Hence, if we construct the similar system as (1), the previous result implies that there exists a finite subsystem which has exactly the same solutions and Makanin's result is again applicable. However, the finite subsystem is not effective here either. Nor do we have any obvious subclasses for which the construction is effective since it is possible that the subsystem is not the same as the test set for the given language.

Also for the prime languages defined by equality method the argument in Theorem 2.7 is not sufficient. If the system of equations fails to have a solution then the language cannot be a prime language in this sense either; otherwise, we do not know whether the language under examination contains all or only some of the prime solutions of $P(g,h)$, $S(g,h)$ or $F(g,h)$.

# 3 Binary case

In this section we consider prime words and prime languages in a binary alphabet. It seems that in many cases the results form an exception compared with larger alphabets. Also the instances seem to be much more limited even though we lack the exact evidence. Hence this section concentrates more on conjectures than on actual results.

The first conjecture deals with P-words.

**Conjecture 3.1** *In a binary alphabet $w$ is a P-word if and only if*

$$\psi(w) = \psi_0(w) \quad \text{or} \quad w \in a^+b^+ \cup b^+a^+. \tag{2}$$

It actually looks like any word which does not satisfy (2) is periodicity forcing. If this could be proved, Conjecture 3.1 would be a straight consequence of Lemma 2.6.

Another interesting thing is the hierarchy of P-words and S-words. In [4, 5] we proved that in alphabets with at least three letters the inclusion is strict; still, in a binary alphabet they seem to be equal.

**Conjecture 3.2** *Let $g$ and $h$ be morphisms over a binary alphabet. Then*

$$P(g,h) = S(g,h).$$

Here it would suffice to show that there are no other equality sets generated by two words, apart from the sets $\{a,b\}$ and $\{a^ib, ba^i\}$, $i \geq 1$.

This is closely connected also with the following conjecture of prime languages (with at least two words) defined in equality sense.

**Conjecture 3.3** *In the binary alphabet $\{a,b\}$ the only P- and S-languages are*

1. $\{a^ib, ba^i\}$, $i \geq 1$, *and*

2. $c(\{a^ib^j\}(i,j)=1)$

*and the only F-languages, in addition to 1.,*

2′. $\{w \in \Sigma^+ \mid \psi_0(w) = (i,j)$ *and $w$ is ratioprimitive*$\}$ *for some* $i,j \geq 1$.

Here the set $c(\{a^ib^j\}(i,j)=1)$ consists of all the words which are permutations of $a^ib^j$ where $\gcd(i,j)=1$. For instance, $c(\{a^1b^3\}) = \{abbb, babb, bbab, bbba\}$.

For prime languages with at least three letters we have better results.

**Theorem 3.4** *The properties of being a P- or an S-language are both decidable for languages with cardinality at least three over a binary alphabet.*

**Theorem 3.5** *If $L$ is a finite language with cardinality at least three in a binary alphabet then $L$ is not an F-language.*

The first result concerns also prime languages defined in inclusion sense but not the second one. In fact the inclusion definition carries much more information about prime languages.

**Theorem 3.6** *$L \subset \{a,b\}^*$ is an F-language (in inclusion sense) if and only if its members are ratioprimitive and have the same basic Parikh vector.*

**Proof:** The "only if"-part follows from Theorem 2.3 and the fact that in a binary alphabet all the members of the equality set must have the same basic Parikh vector. On the other hand, if the words of $L$ have the same basic Parikh vector then they must be solutions for some periodic instance $(g,h)$ whereas ratioprimitiveness now guarantees that they are, indeed, F-prime solutions. By Lemma 2.2 any such instance $(g,h)$ can be effectively constructed. □

For S- and P-languages the following theorem expresses a sufficient condition but we conjecture that the condition is also necessary.

**Theorem 3.7** *$L \subset \{a,b\}^*$ is a P- and an S-language (in inclusion sense) if either its members have the same basic Parikh vector and for each $w \in L$, $\psi(w) = \psi_0(w)$ or $L \subset \{a^ib, ba^i\}$ (or symmetrically $\{ab^i, b^ia\}$), $i \geq 1$.*

In larger alphabets, however, we lack any similar knowledge of equality sets. Thus the only extension of Theorem 3.6 we can prove is that a given language $L$ is an F-language only if its members are ratioprimitive.

# References


[1] K. Culik II, J. Karhumäki: On the equality sets for homomorphisms on free monoids with two generators, *RAIRO Inform. Théor.* **14** (1980) 349–369.

[2] K. Culik II, J. Karhumäki: Systems of equations over a free monoid and Ehrenfeucht's Conjecture, *Discrete Math.* **43** (1983) 139–153.

[3] M. Lipponen: Primitive words and languages associated to PCP, *EATCS Bull.* **53** (1994) 217–226.

[4] M. Lipponen: Post Correspondence Problem: words possible as primitive solutions, *Proc. 22nd ICALP, Springer LNCS* **944** (1995) 63–74.

[5] M. Lipponen: On F-prime solutions of the Post Correspondence Problem, *2nd Internat. Conf. on Developments in Language Theory*, Magdeburg, 1995, to appear.

[6] M. Lipponen: On primitive solutions of the Post Correspondence Problem, *TUCS Dissertations* No. 1 (1996).

[7] M. Lipponen, Gh. Păun: Strongly prime PCP words, *Discrete Appl. Math.* **63** (1995) 193–197.

[8] G.S. Makanin: The problem of solvability of equations in a free semigroup (in Russian), *Mat. Sb.* **103** No. 145 (1977) 148–236.

[9] A. Mateescu, A. Salomaa: PCP-prime words and primality types, *RAIRO Inform. Théor.* **27** (1993) 57–70.

[10] E. Post: A variant of a recursively unsolvable problem, *Bull. Amer. Math. Soc.* **53** (1946) 264–268.

[11] G. Rozenberg, A. Salomaa: *Cornerstones of Undecidability*, Prentice Hall (1994).

[12] G. Rozenberg, A. Salomaa (ed.): *Handbook of Formal Languages, I–III*, Springer-Verlag, forthcoming.

[13] A. Salomaa: The Ehrenfeucht conjecture: a proof for language theorists, *EATCS Bull.* **27** (1985) 71–82.

[14] A. Salomaa, K. Salomaa, Sheng Yu: Primality types of instances of the Post Correspondence Problem, *EATCS Bull.* **44** (1991) 226–241.

**Turku Centre for Computer Science**
**Lemminkäisenkatu 14**
**FIN-20520 Turku**
**Finland**

http://www.tucs.abo.fi

**University of Turku**
- **Department of Mathematical Sciences**

**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**