(G)PCP for words of length at most two

Vesa Halava
Tero Harju

Department of Mathematics, University of Turku
FIN-20014 Turku, Finland
and TUCS - Turku Centre for Computer Science

Mika Hirvensalo
Juhani Karhumäki

Department of Mathematics, University of Turku
FIN-20014 Turku, Finland
and TUCS - Turku Centre for Computer Science

Supported by the Academy of Finland under grant 44087.

E-mail: `{vehalava,harju,mikhirve}@utu.fi`
`karhumak@cs.utu.fi`

Turku Centre for Computer Science
TUCS Technical Report No 463
June 2002
ISBN 952-12-1013-3
ISSN 1239-1891

Abstract

We prove that the generalized Post Correspondence Problem is undecidable in the case where the lengths of the words are at most two and the size (or the number of pairs of words) is 32. The proof uses the undecidability result of Tzeitin, namely the undecidability of the word problem of a semigroup. We also transform our constructions in order to achieve a proof for the undecidability of the Post Correspondence Problem with images of length 1 or 2.

Keywords: Post Correspondence Problem, generalized Post Correspondence Problem,undecidability

TUCS Laboratory
Automata Theory and Combinatorics on Words

1 Introduction

Let A and B be two finite alphabets and h, g be two morphisms $h, g\colon A^* \to B^*$. The *Post Correspondence Problem*, PCP for short, is to determine if there exists a nonempty word $w \in A^*$ such that $h(w) = g(w)$. It was proved by Post [8] that this problem is undecidable in general. Such a word w that $h(w) = g(w)$ is called a *solution* of the *instance* (h, g) of the PCP. The *size* of the instance (h, g) is the size of the domain alphabet A.

Another important problem is the *generalized* PCP, GPCP for short. It consists of two morphisms $h, g\colon A^* \to B^*$ and words $p_1, p_2, s_1, s_2 \in B^*$. The GPCP is to tell whether or not there exists a nonempty word $w \in A^*$ such that

$$p_1 h(w) s_1 = p_2 g(w) s_2.$$

Here again w is called a *solution*. We shall denote the instance of the GPCP by $((p_1, p_2), h, g, (s_1, s_2))$. The pair (p_1, p_2) is called the *begin words* and (s_1, s_2) is called the *end words*. Note that the GPCP is also undecidable, the reason being that the PCP is a subproblem of it.

The PCP is very useful in proving other undecidability results, not only in combinatorics on words or language theory, but also in other fields of mathematics, for example in the theory of matrices, see [4] or [2].

Also, the boundary between the decidabilty and undecidabilty can be investigated by restricting the morphisms. For example it is known that the binary PCP is decidable, i.e., the PCP is decidable for instances of size $|A| \leq 2$, see [1] and [3], and it is undecidable, if $|A| \geq 7$, see [7]. The same bounds hold for the GPCP, see [1] and [3] for the binary case and [5] for the undecidability of the GPCP for $|A| = 7$. For the PCP and GPCP, the decidability status is open for the sizes between these two bounds.

In this paper we shall restrict the lengths of the images. Note that it is clear that if the images are all of length one, then the PCP and the GPCP are trivially decidable. We shall prove that both PCP and GPCP are undecidable in the case where the lengths of the images are 1 or 2. This problem is an exercise in Hopcroft and Ullman [6]. We do not know whether any proof for this result is ever published. We assume that an undecidability proof using the halting problem for Turing machines is presumed in Hopcroft's and Ullman's solution. But we shall present a nice proof, which uses the undecidability of the word problem of the semigroup S_7 defined by Tzeitin in [9].

2 Tzeitin semigroup

Our proof of the undecidability of the PCP (and GPCP) for case where the images are of length 1 or 2 uses the undecidability of the word problem in the semigroup S_7 defined by Tzeitin [9]. Consider the following semigroup S_7 with a set of five generators $\Gamma = \{a, b, c, d, e\}$ and seven relations:

$S_7 = \langle a, b, c, d, e \mid R \rangle$,
$R = \{ac = ca, ad = da, bc = cb, bd = db, eca = ce, edb = de, cca = ccae\}$.

Here the relations in R are the 'rewriting rules' of the semigroup.

The problem whether $u = v$ in a (presentation of a) semigroup $S = \langle X \mid R \rangle$ is called the *word problem of* S. Recall that $u = v$ in S if there is a sequence u_i, $i = 0, 1, \ldots, n$, such that $u_0 = u$, $u_n = v$ and for $i = 0, 1, \ldots n-1$,

$$u_i = x_i \alpha_i y_i, \quad u_{i+1} = x_i \beta_i y_i,$$

where $(\alpha_i = \beta_i) \in R$. This will be denoted by $u = v \in S$.

Tzeitin [9] proved the following theorem.

Theorem 1. *It is undecidable, for words $u, v \in \Gamma^*$, whether or not $u = v$ in S_7.*

We shall next define the semigroup S_7', which is used in the proof of our main result. We define few new letters $\langle xy \rangle$, where $x, y \in \Gamma$. Now the generator set of S_7' is

$$\Gamma' = \Gamma \cup \{\langle ae \rangle, \langle cc \rangle, \langle ed \rangle, \langle ec \rangle\}$$

and the set of relations is

$$\begin{aligned} R' = \{ & ac = ca, ad = da, bc = cb, bd = db, \\ & \langle ec \rangle a = ce, \langle ed \rangle b = de, \langle cc \rangle a = \langle cc \rangle \langle ae \rangle\} \\ & \cup \{\langle ae \rangle = ae, \langle cc \rangle = cc, \langle ed \rangle = ed, \langle ec \rangle = ec\}. \end{aligned}$$

The following theorem is obvious by the definition of S_7' and Theorem 1.

Theorem 2. *It is undecidable, for words $u, v \in \Gamma^*$, whether or not $u = v$ in S_7'.*

3 GPCP

In this section we shall prove our main result, that is

Theorem 3. *The GPCP is undecidable for the instances where the lengths of the images of the morphisms are at most 2 and the size is 32.*

	$\#$	x	$[s=t]$
g	$\#$	x	t
h	$\#$	x	s

Table 1: Definitions of g and h, where $x \in \Gamma'$ and $s = t \in R'$

Define the domain alphabet as

$$A = \Gamma' \cup \{\#\} \cup \{[s=t] \mid [s=t] \in R'\}.$$

Note that the relations of R' are symmetric, therefore there we define a symbol also for the reverse relations of the set R'. For example, both $[ac = ca]$ and $[ca = ac]$ are in A.

The morphisms g and h (from A^* to $(\Gamma' \cup \#)^*$) are defined in Table 1.

For $u, v \in \Gamma$, we define the instance of the GPCP to be

$$I_{u,v} = ((\#u\#, \#), g, h, (\#, \#v\#)).$$

Lemma 1. *Let u, v and $I_{u,v}$ be as above. Then $u = v \in S_7'$ if and only if $I_{u,v}$ has a solution.*

Proof. Suppose $u = v \in S_7'$. Then there is a sequence $u = u_1, u_2, \ldots, u_k = v$, where $u_i = u_i'\alpha_i u_i''$ and $u_{i+1} = u_i'\beta_i u_i''$, and $\alpha_i = \beta_i \in R'$. We construct a solution to $I_{u,v}$ by induction on k.

If $k = 1$, then $u = v$ already in Γ^+. In this case u is a solution to $I_{u,v}$.

Assume now that $k \geq 2$. Let $I'_{u,u_{k-1}} = ((\#u\#, \#), g, h, (\#, \#u_{k-1}\#))$ be the instance of the GPCP corresponding to $u = u_{k-1} \in S_7'$. By the induction hypothesis, $I'_{u,u_{k-1}}$ has a minimal length solution x'. Then if $\alpha_{k-1} = \beta_{k-1} \in R'$ and $r = [\alpha_{k-1} = \beta_{k-1}]$, then $x = x'\#u'_{k-1}ru''_{k-1}$ is a solution of $I_{u,v}$, since $\#u\#g(x')\# = \#h(x')\#u_{k-1}\#$, and so

$$\#u\#g(x)\# = \#u\#g(x')\#v\# = \#h(x)\#v\# = \#h(x')\#u_{k-1}\#v\#.$$

Suppose $I_{u,v}$ has a solution x. We can assume that x is of minimal length. Clearly, this x must be of the form $x_1\#x_2\#\cdots x_{k-1}\#x_k$, where $x_i \in (A\backslash\#)^*$. Hence $g(x) = g(x_1)\#\cdots g(x_{k-1})\#g(x_k)$ and $h(x) = h(x_1)\#\cdots h(x_{k-1})\# h(x_k)$, and then $h(x_1) = u$ and $g(x_m) = v$. Now $\#h(x)\#v\# = \#u\#g(x)\#$ must be of the form

$$\#u_1\#u_2\#\cdots\#u_{k-1}\#u_k\#,$$

where $u_i \in (\Gamma')^*$, $u_1 = u$ and $u_k = v$. We show that $u_i = u_{i+1} \in S_7'$ for every $1 \leq i \leq k-1$, from which $u = v \in S_7'$ follows.

Because $u = h(x_1)$, there are two possibilities either $x_1 = u$ or $x_1 = w_1 r_{i_1} w_2 r_{i_2} \cdots w_l$, with $w_i \in (\Gamma')^*$ and $r_i = [s_i = t_i]$. The first case yields that also $h(x_1) = u$ and necessarily $u = v$, since x is the minimal solution. Therefore we may assume that $x_1 = w_1 r_{i_1} w_2 r_{i_2} \cdots w_l$. Then $u = h(w_1 r_{i_1} w_2 r_{i_2} \cdots w_l) = w_1 s_{i_1} w_2 s_{i_2} \cdots w_l$, and $u_2 = g(x_1) = g(w_1 r_{i_1} w_2 r_{i_2} \cdots w_l)$ $= w_1 t_{i_1} w_2 t_{i_2} \cdots w_l$. Therefore $u_1 = u_2 \in S_7'$, because $u_1 (= u)$ and u_2 only differ by u_2 having t_i where u_1 has s_i.

Continuing this reasoning, we can show that for every two words $u_i, u_{i+1} \in \Gamma^*$ occurring in

$$\#h(x)\#v\# = \#u_1\#u_2\# \cdots \#u_{k-1}\#u_k\# = \#u\#g(x)\#$$

we have $u_i = u_{i+1} \in S_7'$. Hence $u = v \in S_7'$, since $h(x)$ starts with $u_1 = u$ and $g(x)$ ends with $u_k = v$. $\square$

Note that $|A| = 9 + 22 + 1 = 32$. Together with Tzeitin's result, the lemma above implies Theorem 3.

4 PCP

Next we shall transform the construction for the undecidability of the GPCP to the PCP and give a proof for the undecidability of the PCP for words of length 1 or 2. This problem is an exercise in Hopcroft and Ullman [6]. We assume that an undecidability proof using the halting problem for Turing machines is presumed in their solution. We shall here present a proof which uses the word problem and the transformation S_7' of the Tzeitin semigroup. It also seems that the same ideas should be used to reduce an instance of the halting problem to an instance of the PCP with lengths of images 1 or two. Note that in the undecidability proof of the PCP, the halting problem is reduced to the *modified PCP* and then to the PCP. This cannot be done directly when proving the undecidability of the PCP for words of length 1 or 2, since in the reduction of the modified PCP to the PCP the lengths of the images are doubled.

We will force the instance to write to words u and v as a concatenation of images and therefore the alphabet size is no longer bounded. We also need to take copies of the letters exclude the trivial solutions.

First of all, let

$$A' = \Gamma' \cup \underline{\Gamma'} \cup \{F, E, \#, \underline{\#}, r_1, r_2, \ldots, r_{11}, r_1', r_2', \ldots r_{11}'\} \cup U,$$

where $\underline{\Gamma'}$ is underlined copy of Γ', $r_1, \ldots, r_{11}$ are the rules of R' and $r_1', \ldots, r_{11}'$ new copies of them and U a set of letters defined for the words u and w. Note that we do not have symbols for reverse relations of R' in this case, only the primed copies of the 11 relations.

Clearly, the only possible beginning of a solution is F. Now we define the set U, which takes care of forcing the solution begin so that u is written by g and end so that h writes v to the end.

	F	E	$\#$	$\underline{\#}$	x	$\underline{x}$	$[s=t]$	$[s=t]'$
g	Fa_1	E	$\underline{\#}$	$\#$	$\underline{x}$	x	$\underline{t}$	s
h	F	$b_\ell E$	$\#$	$\underline{\#}$	x	$\underline{x}$	s	$\underline{t}$

Table 2: Definitions of g and h, where $x \in \Gamma'$ and $s = t \in R'$

Assume $|u| = k$ and $|v| = \ell$ and let $U = \{a_1 \ldots, a_k, b_1, \ldots, b_\ell\}$. The images of a_i's are of length 2 for g and of length 1 for h and the images of b_i's are of length 1 for g and of length 2 for h. They are defined so that

$$g(Fa_1 \cdots a_k) = Fa_1a_2 \cdots a_k u\#, \quad h(Fa_1 \cdots a_k) = Fa_1 \cdots a_k$$
$$h(b_1 \cdots b_\ell E) = \#vb_1b_2 \cdots b_\ell E, \quad g(b_1 \cdots b_\ell E) = b_1b_2 \cdots b_\ell E.$$

Theorem 4. *The PCP is undecidable for instances where the lengths of the images are one or two.*

Proof. The proof goes as the proof of Theorem 3. Indeed, we can proved that for words $u, v \in \Gamma^*$, $u = v \in S_7'$ if and only if there exists a word $x = Fa_1 \cdots a_k w b_1 \cdots b_\ell E$ such that

$$\begin{aligned} h(x) &= Fa_1 \cdots a_k u_1 \# \underline{u_2} \# u_3 \# \cdots \# u_{k-1} \# \underline{u_k} b_1 \cdots b_\ell E \\ &= g(x) \end{aligned}$$

and $u_1 = u$, $u_k = v$ and $u_i = u_{i+1} \in S_7'$ for all $1 \leq i \leq k-1$. Note that every other word u_i is underlined, but the proof goes as the proof of Theorem 3 if we ignore the underlining. Note also that we do not have rewriting rules for reverses of R', but the reverse rewritings are done with the primed copies of rules of R'. □

Note that using the similar ideas one could produce a proof for Theorems 3 and 4 by using the undecidability of the halting problem for the Turing machines. But the spliting of the rules of Turing machines into rules using only two symbols is more difficult than splitting the rewriting rules of Tzeitin's semigroup.

References

[1] A. Ehrenfeucht, J. Karhumäki, and G. Rozenberg. The (generalized) Post Correspondence Problem with lists consisting of two words is decidable. *Theoret. Comput. Sci.*, 21:119–144, 1982.

[2] V. Halava. Decidable and undecidable problems in matrix theory. Master's thesis, University of Turku. TUCS Tech. Report **127**, 1997.

[3] V. Halava, T. Harju, and M. Hirvensalo. Binary (generalized) Post Correspondence Problem. *Theoret. Comput. Sci.*, 276(1-2):183–204, 2002.

[4] T. Harju and J. Karhumäki. Morphisms. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 1. pp. 439–510, Springer-Verlag, 1997.

[5] T. Harju, J. Karhumäki, and D. Krob. Remarks on generalized Post Correspondence Problem. *Lecture Notes in Comput. Sci.* 1046, pages 39–48. Springer-Verlag, 1996.

[6] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley, 1979.

[7] Y. Matiyasevich and G. Sénizergues. Decision problems for semi-Thue systems with a few rules. In *Proceedings, 11$^{\text{th}}$ Annual IEEE Symposium on Logic in Computer Science*, pages 523–531, New Brunswick, New Jersey, 27–30 July 1996. IEEE Computer Society Press.

[8] E. Post. A variant of a recursively unsolvable problem. *Bull. of Amer. Math. Soc.*, 52:264–268, 1946.

[9] G. C. Tzeitin. Associative calculus with an unsolvable equivalence problem. *Tr. Mat. Inst. Akad. Nauk*, (52):172–189, 1958. (In Russian)

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

http://www.tucs.fi

University of Turku
- **Department of Information Technology**
- **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**