Solving equations under Simon's congruence

Péter Pál Pach*

Department of Computer Science and
Information Theory
Budapest University of Technology and
Economics
1521 Budapest, Magyar tudósok körútja 2.,
Hungary
ppp@cs.bme.hu

Abstract: Simon's congruence, denoted by $\sim_k$, relates the words having the same subwords of length at most k. In this paper a normal form is presented for the equivalence classes of $\sim_4$. Moreover, a canonical solution of the equation $pwq \sim_2 r$ is also shown (p, q, r are the parameters), which can be viewed as a generalization of giving a normal form for $\sim_2$.

Keywords: combinatorics of words, normal form, piecewise testable languages

1 Introduction

The theory of formal languages goes back to natural languages. Linguists, e.g. Chomsky, gave mathematical definitions of natural concepts such as words, languages and grammars: Given a finite set A, a word on A is simply an element of the free monoid on A, and a language is a set of words. This theory connects languages, automata and semigroups.

One of the bases of formal language theory is Kleene's theorem: It states that the class of recognizable languages (e.g. recognized by finite automata) coincides with the class of rational languages, which are given by rational expressions. Rational expressions are the generalization of polynomials involving three operations: union, product and star operation. As another crucial point, Schützenberger showed that there is an equivalence between finite automata and finite semigroups. He proved that a finite monoid, the so-called syntactic monoid, can be assigned to each recognizable language; this is the smallest monoid recognizing the language.

A large class of languages is the family of piecewise testable languages, which has been deeply studied in formal language theory, see for example, Simon [6] or Stern [7]. Formally, a language L is k-piecewise testable, if $x \in L$ and $x \sim_k y$ implies that $y \in L$, where $x \sim_k y$ if and only if x and y have the same subwords of length at most k. It is easy to see that $\sim_k$ is a congruence, the so-called Simon's congruence, with finite index. Some estimations of this index can be found in [3] and [4]. Furthermore, in [4] the word problem for the syntactic monoids of the varieties of k-piecewise testable languages are analyzed and a normal form of the words is presented for $k = 2$ and 3. In this paper our aim is to give a normal form when $k = 4$. The new idea is to investigate a more general question, namely, to determine a canonical solution of the equation $pwq \sim_k r$. It is going to be seen that if a canonical solution of the equations of the form $pwq \sim_k r$ can be defined for some k, then a normal form can be defined for $k + 2$.

2 Preliminaries

At first, some basic notions and definitions are going to be introduced. The word w is a *subword* of u, if w is a sequence of not necessarily consecutive variables taken from u. Given an integer $k > 0$, let $u \sim_k v$

*Research is supported by the OTKA research grant K108947.

if and only if the words u, v have the same set of subwords of length at most k. A language L over an alphabet X is *k-piecewise testable* if and only if L is a union of classes of the equivalence relation $\sim_k$. Another characterization says that a language L over an alphabet X is k-piecewise testable if and only if it is a finite boolean combination of languages of the form

$$X^* x_1 X^* x_2 X^* \dots X^* x_l X^*, \text{ where } x_1, \dots, x_l \in X,\ 0 \leq l \leq k.$$

A language is piecewise testable, if there exists a natural number k such that the language is k-piecewise testable.

Simon [6] found a basis of identities for k-piecewise testable languages, if $k = 1, 2$. Moreover, Blanchet-Sadri [1, 2] gave a basis of identities for $k = 3$, and proved that there is no finite basis of identities for $k > 3$. If one is interested in the basic definitions and theorems in more detail, they can read about them in Pin [5].

In this paper the alphabet X is going to be an n-element set (for some $n \in \mathbb{N}$), namely $X = \{x_1, x_2, \dots, x_n\}$. For a word w let us denote the set of its subwords of length at most k by $(w)_k$. This way $w_1 \sim_k w_2$ if and only if $(w_1)_k = (w_2)_k$, thus we can refer to the $\sim_k$-equivalence class of a word w by $(w)_k$. The set of the 1-element subwords of w is the content of w, let us denote it by $c(w)$. Clearly, $(w)_k$ determines $c(w)$.

3 Solving the equation $pwq \sim_2 r$

In this section our aim is to define a *canonical solution* of the equation $pwq \sim_2 r$, or equivalently $(pwq)_2 = (r)_2$, if a solution exists. Here the words p, q, r are parameters, and we would like to solve the equation for w. By the term *canonical solution* we mean that the solution should only depend on the equivalence classes of the words p, q, r, that is, on $(p)_2$, $(q)_2$ and $(r)_2$. This approach is a generalization of finding a normal form for the words under $\sim_2$ (that is, a normal form for the elements of the free syntactic monoid of the 2-piecewise testable languages). Namely, the normal form of the word r can be defined as the canonical solution of the equation $(pwq)_2 = (r)_2$ when we set p and q to be the empty word.

Assume that the words p, q, r are given, and our aim is to find an above mentioned well-defined solution $\overline{w}^{(p,q,r)} = \overline{w}$ of the equation $(pwq)_2 = (r)_2$. At first some observations are made about the set which contains the subwords of w having length at most 2: $(w)_2$. Let us define $\mathcal{A}$ and $\mathcal{B}$ as follows:

$$\mathcal{A} = \{u_1u_2, u_1, u_2 \mid u_1u_2 \in (r)_2, u_1 \notin c(p), u_2 \notin c(q)\} \cup \{\textit{empty word}\} \cup$$
$$\cup \{u_1 \mid u_1 \notin c(p) \text{ and } \exists u_2 : u_1u_2 \in (r)_2, u_1u_2 \notin (q)_2\} \cup \{u_2 \mid u_2 \notin c(q) \text{ and } \exists u_1 : u_1u_2 \in (r)_2, u_1u_2 \notin (p)_2\},$$

$$\mathcal{B} = \{u_1u_2 \mid u_1u_2 \notin (r)_2\} \cup \{\textit{empty word} \mid (pq)_2 \setminus (r)_2 \neq \emptyset\} \cup$$
$$\cup \{u_1 \mid \exists u_2 \in c(q) : u_1u_2 \notin (r)_2\} \cup \{u_2 \mid \exists u_1 \in c(p) : u_1u_2 \notin (r)_2\}.$$

(Here u_1 and u_2 denote single letters.)
The following statements can be easily checked:

Proposition 1 *For the above defined sets $\mathcal{A}$ and $\mathcal{B}$ we have that:*

- $(pwq)_2 \supseteq (r)_2$ *if and only if* $\mathcal{A} \subseteq (w)_2$,
- $(pwq)_2 \subseteq (r)_2$ *if and only if* $\mathcal{B} \cap (w)_2 = \emptyset$.

For example, if $p = x_1,\ q = x_2,\ r = x_2x_1x_1x_3x_2$, then our equation is

$$(x_1wx_2)_2 = (x_2x_1x_1x_3x_2)_2,$$

and we have

$$\mathcal{A} = \{x_1, x_2, x_3, x_2x_1, x_2x_3\},\ \mathcal{B} = \{x_3x_1, x_3x_3\}.$$

The word $w = x_2x_1x_3$ is a solution of this equation, since both $\mathcal{A} \subseteq (w)_2$ and $\mathcal{B} \cap (w)_2$ hold.

Note that the sets $\mathcal{A}$ and $\mathcal{B}$ only depend on the $\sim_2$-equivalence classes of p, q and r, moreover, $\mathcal{A} \cup \mathcal{B}$ may not contain all the words of length at most 2. Accordingly, we obtained as a key observation that a word w satisfies the equation $(pwq)_2 = (r)_2$ if and only if $\mathcal{A} \subseteq (w)_2$ and $(w)_2 \cap \mathcal{B} = \emptyset$.

Clearly, if $\mathcal{A} \cap \mathcal{B} \neq \emptyset$, then the equation has no solution. However, $\mathcal{A} \cap \mathcal{B}$ does not straightforwardly yield that there is a solution, since for arbitrary sets $\mathcal{A}$ and $\mathcal{B}$ containing words of length at most 2 and satisfying $\mathcal{A} \cap \mathcal{B} = \emptyset$, it is possible that there is no word w such that $\mathcal{A} \subseteq (w)_2$ and $\mathcal{B} \cap (w)_2 = \emptyset$ (even if $\mathcal{A}$ is downward closed and $\mathcal{B}$ is upward closed). For instance, if $\mathcal{A} = \{x_1, x_2, \textit{empty word}\}$ and $\mathcal{B} = \{x_1x_2, x_2x_1\}$, then there is no w such that $\mathcal{A} \subseteq (w)_2$ and $(w)_2 \cap \mathcal{B} = \emptyset$.

Let us try to find the word $\overline{w}$ satisfying the conditions $\mathcal{A} \subseteq (\overline{w})_2$ and $(\overline{w})_2 \cap \mathcal{B} = \emptyset$, moreover containing each variable at most twice. Now we define a directed graph $G = (V, E)$. The vertices of the graph correspond to the variables in $c(\overline{w})$: y_i represents the first appearance of the variable x_i and z_i the last appearance of it. (The alphabet is $\{x_1, x_2, \ldots, x_n\}$.) We choose V in such a way that it satisfies the following conditions:

- If $x_i \in \mathcal{A}$ and $x_ix_i \in \mathcal{B}$, then let $y_i = z_i \in V$.
- If $x_i \in \mathcal{A}$ and $x_ix_i \notin \mathcal{B}$, then let $y_i, z_i \in V$, $y_i \neq z_i$.
- If $x_i \notin \mathcal{A}$, then $y_i, z_i \notin V$.
- If $i \neq j$, then $y_i \neq y_j, z_i \neq z_j, y_i \neq z_j$.

Therefore, the set of the vertices of the graph is

$$V = \{y_i \mid x_i \in \mathcal{A}, x_ix_i \in \mathcal{B}\} \cup \{y_i, z_i \mid x_i \in \mathcal{A}, x_ix_i \notin \mathcal{B}\}.$$

A directed edge from a vertex u to another vertex v represents that u must appear before v in $\overline{w}$. For instance, if $y_1 \to z_2$ is a directed edge, then in $\overline{w}$ the first appearance of x_1 must preceeds the last appearance of x_2. The edges of G are obtained in the following way:

(i) If $x_ix_j \in \mathcal{A}$ (where $i \neq j$), then let $y_iz_j \in E$.

(ii) If $z_j, y_i \in V$ and $x_ix_j \in \mathcal{B}$ (where $i \neq j$), then let $z_jy_i \in E$.

(iii) If $y_i \in V$ and $y_i \neq z_i$, then let $y_iz_i \in E$.

Hence, the set of the edges of the graph is

$$E = \{y_iz_j \mid x_ix_j \in \mathcal{A}\} \cup \{z_jy_i \mid z_j, y_i \in V \text{ and } x_ix_j \in \mathcal{B}\} \cup \{y_iz_i \mid x_i \in \mathcal{A}, x_ix_i \notin \mathcal{B}\}.$$

The following proposition gives a characterisation of the solvability of the equation $(pwq)_2 = (r)_2$:

Proposition 2 *The equation $(pwq)_2 = (r)_2$ is solvable if and only if $\mathcal{A} \cap \mathcal{B} = \emptyset$ and there is no directed cycle in the graph G.*

Proof: At first assume that the equation has a solution, let's denote it by $\overline{w}$. Since $\mathcal{A} \subseteq (\overline{w})_2$ and $(\overline{w})_2 \cap \mathcal{B} = \emptyset$, we have $\mathcal{A} \cap \mathcal{B} = \emptyset$. If $x_ix_i \in \mathcal{A}$, then x_i has to appear in $\overline{w}$ at least twice. If $x_i \in \mathcal{A}$ and $x_ix_i \notin \mathcal{B}$, then x_i has to appear in $\overline{w}$, and without the loss of generality it can be assumed that x_i appears at least twice in $\overline{w}$. Since, if $\overline{w}$ satisfies the equation and contains x_i only once, then if we double x_i (that is, right after the unique appearance of the letter x_i we write x_i again) and obtain the word $\overline{w}^*$, then clearly $(\overline{w})_2 \subseteq (\overline{w}^*)_2 \subseteq (\overline{w})_2 \cup \{x_ix_i\}$, therefore $\overline{w}^*$ is a solution, as well. Hence, it can be assumed that all the letters x_i for which $x_i \in \mathcal{A}$ and $x_ix_i \notin \mathcal{B}$ appear in $\overline{w}$ at least twice. If we delete all except the first and last appearances of every variable, $(\overline{w})_2$ does not change, so it can be assumed that $\overline{w}$ contains each variable at most twice. Then the word $\overline{w}$ can be viewed as a permutation of the vertices of G: The vertex y_i is represented by the first appearance of x_i, the vertex z_i is represented by the last (second) appearance of x_i, and when $x_i \in \mathcal{A}$ and $x_ix_i \in \mathcal{B}$, the vertex $y_i = z_i$ is represented by the unique appearance of x_i. If $y_iz_j \in E$, then $x_ix_j \in \mathcal{A}$, hence in $\overline{w}$ the first appearance of x_i, that is, y_i has to appear before the last appearance of x_j, that is, z_j. If $z_jy_i \in E$, then $x_ix_j \in \mathcal{B}$, so in $\overline{w}$ the last appearance of x_j, that is, z_j has to appear before the first appearance of x_i, that is, y_i. Finally, if $x_i \in \mathcal{A}$ and $x_ix_i \notin \mathcal{B}$, then the first appearance of x_i, that is, y_i has to appear before the last appearance of x_i, that is, z_i, naturally. To sum up, for all edges $uv \in E$ the occurence of the letter corresponding to u must preceed the occurence of the letter corresponding to v in $\overline{w}$. Therefore, G can not contain a directed cycle.

Now assume that $\mathcal{A} \cap \mathcal{B} = \emptyset$ and G does not contain a directed cycle. Then G has a topological ordering, that is, an ordering of the vertices $v_1, v_2, \ldots, v_{|V|}$ satisfying that for all the edges v_iv_j we have $i < j$. Let $\overline{w}$ be the word obtained in the following way: In $v_1v_2 \ldots v_{|V|}$ replace each y_i and z_i by x_i. We claim that $\overline{w}$ is a solution of the equation $(pwq)_2 = (r)_2$. Accoring to Proposition 1 we have to show that each element of $\mathcal{A}$ is in $(\overline{w})_2$ and none of the elements of $\mathcal{B}$ appears in $(\overline{w})_2$. By the definition of the vertex set of G it can be easily seen that the content of $\overline{w}$ is precisely the set of the letters occuring as a 1-length subword in $\mathcal{A}$. Then the condition $\mathcal{A} \cap \mathcal{B} = \emptyset$ implies that the 1-length words in $\mathcal{B}$ are not in $(\overline{w})_2$. Now, it remains to check the 2-length words. If $x_ix_i \in \mathcal{A}$, then x_i appears twice in $\overline{w}$, therefore $x_ix_i \in (\overline{w})_2$. If $x_ix_j \in \mathcal{A}$ for some $i \neq j$, then $y_iz_j \in E$, hence the first appearance of x_i is before the last appearance of x_j in $\overline{w}$, so $x_ix_j \in (\overline{w})_2$. Therefore, we obtained that $\mathcal{A} \subseteq (\overline{w})_2$. Now, if $x_ix_i \in \mathcal{B}$, then x_i appears at most once in $\overline{w}$, so $x_ix_i \notin (\overline{w})_2$. If $x_ix_j \in \mathcal{B}$ for some $i \neq j$, then $z_jy_i \in E$, so in $\overline{w}$ the last appearance of x_j is before the first appearance of x_i, thus $x_ix_j \notin (\overline{w})_2$. Therefore, $\mathcal{B} \cap (\overline{w})_2 = \emptyset$, hence $\overline{w}$ is indeed a solution. $\square$

To sum up, we have seen that the graph G is well-defined by $(p)_2, (q)_2, (r)_2$, the sets $\mathcal{A}$ and $\mathcal{B}$ are also well-defined and, of course, we can choose a topological ordering of the vertices of G in a canonical way, if G is acyclic. For instance, the one achieved by running depth first search and taking the vertices in reverse order respect to their finishing times. Accordingly, a canonical solution $\overline{w} = \overline{w}^{(p,q,r)}$ is obtained this way. Note that the length of $\overline{w}$ is at most $2n$, where n is the size of the alphabet.

The characterization in Proposition 2 can be formulated in the following way as well.

Proposition 3 *The condition $\mathcal{A} \cap \mathcal{B} = \emptyset$ implies that G is a directed acyclic graph, so $\mathcal{A} \cap \mathcal{B} = \emptyset$ is a necessary and sufficient condition for the solvability of the equation $(pwq)_2 = (r)_2$.*

PROOF: In order to prove this let $\hat{r}$ be the $\sim_2$-normal form of r and w^* be the word obtained from $\hat{r}$ after deleting the letters occuring as a 1-length word in $\mathcal{B}$. We claim that $(pw^*q)_2 = (r)_2$. At first, $\mathcal{A} \subseteq (r)_2$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$ implies that $\mathcal{A} \subseteq w^*$. The 1-length words of $\mathcal{B}$ do not occur in w^* (since we deleted them from $\hat{r}$), moreover the 2-length words of $\mathcal{B}$ neither, since they do not occur in $(\hat{r})_2 = (r)_2$ and $(r)_2 \supseteq (w^*)_2$. $\square$

Actually, this construction provides us another possible way to define a canonical solution w^* of the equation $(pwq)_2 = (r)_2$. However, this construction uses the normal form for $\sim_2$, while the previously described one does not use the $\sim_2$-normal form.

Finally we summarize the result of this section in the following proposition:

Proposition 4 *Let p, q, r, p', q', r' be words and suppose that the equation $pwq \sim_2 r$ has a solution. Then $p\overline{w}^{(p,q,r)}q \sim_2 r$. If $p \sim_2 p', q \sim_2 q', r \sim_2 r'$, then $\overline{w}^{(p,q,r)} = \overline{w}^{(p',q',r')}$. Hence, $\overline{w}^{(p,q,r)}$ is a canonical form of a solution of the equation $pwq \sim_2 r$. The length of $\overline{w}^{(p,q,r)}$ is at most $2n$.*

Note that it is possible that $pwq \sim_2 r$, but $w \not\sim_2 \overline{w}^{(p,q,r)}$.

4 Normal form for $k = 4$

In this section our aim is to present a normal form for the words ($\sim_4$-equivalence classes) when $k = 4$. This normal form is going to be given with the help of the canonically defined solution of equations of the form $pwq \sim_2 r$ (where the words p, q, r are parameters). More generally, it is going to be shown that if for some k a canonical solution of the equation $pwq \sim_{k-2} r$ is defined for every p, q, r, then a normal form can be constructed for the words in the case of k. As we defined such a solution in the previous section for $\sim_{4-2}=\sim_2$, this will provide us a normal form for $k = 4$.

Hence, let us assume that a "canonical solution" of the equation $(pwq)_{k-2} = (r)_{k-2}$ (where the words p, q, r are parameters) can be defined (if such w exists). Let us denote this canonical solution by $\overline{w} = \overline{w}^{(p,q,r)}$. The word $\overline{w}^{(p,q,r)}$ is determined by the $\sim_{k-2}$-equivalence classes of p, q, r, that is, by $(p)_{k-2}, (q)_{k-2}$ and $(r)_{k-2}$, and it satisfies the equation $(p\overline{w}^{(p,q,r)}q)_{k-2} = (r)_{k-2}$. Note that $(pw_1q)_{k-2} = (r)_{k-2} = (pw_2q)_{k-2}$ might hold with different $(w_1)_{k-2}$ and $(w_2)_{k-2}$, but we only use that one well-defined solution can be obtained in a canonical way (as we obtained such a solution for $k - 2 = 2$ in the previous section).

Now we show that with the help of this "canonical solution" (for every equation of the form $(pwq)_{k-2} = (r)_{k-2}$) a normal form can be defined for $\sim_k$. Let w be a word and let w' denote the word (obtained from w) in which only the first and last occurences of the variables of w are kept, and the others are deleted. Note that the word $w' = y_1y_2 \dots y_t$, where $y_i \in c(w)$, has length at most $2n$. The word w is separated into $t - 1$ (possibly empty) parts by the letters of w':

$$w = y_1u_1y_2u_2 \dots u_{t-1}y_t.$$

In [4] for $k = 3$ we proved that w' is "almost determined" by $(w)_k$ and defined a "normal form" for w', as well. For general k this normal form for w' can be obtained in the same way. In other words, a word $w' = y_1y_2 \dots y_t$ can be given in such a way that it only depends on $(w)_k$ and there exist words $u_1, u_2, \dots, u_{t-1}$ satisfying $(w)_k = (y_1u_1y_2u_2 \dots u_{t-1}y_t)_k$. Now we show that w' and $(u_1)_{k-2}, \dots, (u_{t-1})_{k-2}$ determine $(w)_k$. Let us suppose that $z = z_1z_2z_3$ is a word of length at most k, where the first letter of z is z_1, the last letter of z is z_3 (and z_2 is a word of length at most $k - 2$). Let y_a be the first appearance of the letter z_1 in w' and y_b be the last appearance of the letter z_3 in w'. If $b \le a$, then $z \notin (w)_k$. If $a < b$, then $z \in (w)_k$ if and only if $z_2 \in (u_ay_{a+1} \dots u_{b-1})_{k-2}$. Therefore, w' and $(u_1)_{k-2}, \dots, (u_{t-1})_{k-2}$ determine $(w)_k$ and our aim is to define $u_1, \dots, u_{t-1}$ in such a way that for every first appearance y_a and last appearance y_b the following holds (we know that an appropriate choice exists):

$$(u_ay_{a+1} \dots u_{b-1})_{k-2} = \{m : y_amy_b \in (w)_k\} =: M_{y_a,y_b}(w). \tag{1}$$

At first we determine an order in which the words $(u_i)_{k-2}$ are going to be defined. For $1 \le i \le t - 1$ let n_i be the total number of first appearances in $\{y_{i+1}, \dots, y_t\}$ and last appearances in $\{y_1, \dots, y_i\}$. We define u_i in increasing order according to n_i. Suppose that for some i, the words u_v for which $n_v < n_i$, are already defined. We show that now u_i can be defined, as well. Let $j \le i$ be maximal such that y_j is a first appearance and $i + 1 \le l$ be minimal such that y_l is a last appearance. Since $y_{j+1}, y_{j+2}, \dots, y_i$ are all last appearances and $y_{i+1}, y_{i+2}, \dots, y_{l-1}$ are all first appearances, $\max(n_j, n_{j+1}, \dots, n_{i-1}, n_{i+1}, n_{i+2}, \dots, n_{l-1}) < n_i$, so $u_j, u_{j+1}, \dots, u_{i-1}, u_{i+1}, u_{i+2}, \dots, u_{l-1}$ are already defined. Let $p = u_jy_{j+1}u_{j+1} \dots y_i$, $q = y_{i+1}u_{i+1} \dots u_{l-1}$ and $(r)_{k-2} = M_{y_j,y_l} = \{m \mid y_jmy_l \in (w)_k\}$. The word u_i has to satisfy the equation $(pu_iq)_{k-2} = (r)_{k-2}$, so let us choose u_i as the canonically defined solution of this equation: $u_i := \overline{u}_i = \overline{u}_i^{(p,q,r)}$. Now, we show that for any appropriate choice of the words u_v, that is, for any choice for which all the equations of the form (1) hold, if we replace u_i by the previously defined $\overline{u}_i$, they will still hold. It means that by setting u_i to be $\overline{u}_i$ we can't make a "mistake".

When we check the equation $M_{y_a,y_b} = (u_ay_{a+1} \dots u_{b-1})_{k-2}$ for some first appearance y_a and last appearance y_b (satisfying $a < b$), then the choice of u_i only plays a role if $a \le i < b$. This yields $a \le j$ and $l \le b$. In the special case when $a = j$ and $l = b$, according to the definition of $\overline{u_i}$ we have $(u_jy_{j+1} \dots u_{l-1})_{k-2} = M_{y_j,y_l}$. Here, M_{y_j,y_l} is determined by $(w)_k$, therefore $(u_jy_{j+1} \dots u_{l-1})_{k-2}$ is also determined by $(w)_k$. Using this observation we obtain that for arbitrary $a \le j$ and $l \le b$ the right hand side of

$$(u_ay_{a+1} \dots u_{b-1})_{k-2} = (u_ay_{a+1} \dots y_j)_{k-2}(u_jy_{j+1} \dots u_{l-1})_{k-2}(y_lu_l \dots u_{b-1})_{k-2},$$

does not depend on the choice of u_i (the only restriction for u_i is that it has to satisfy $(u_jy_{j+1} \dots u_{l-1})_{k-2} = M_{y_j,y_l}$). Hence, we can set $u_i := \overline{u_i}$. Therefore, one by one, the words u_i can be defined with the help of a canonical form of a solution of equations of the form $(puq)_{k-2} = (r)_{k-2}$, and finally the normal form $\hat{w} = y_1\overline{u}_1 y_2 \overline{u}_2 \dots \overline{u}_{t-1} y_t$ is obtained.

We summarize the results of this section in the following proposition:

Proposition 5 *Let v and w be two words. Then $w \sim_4 \hat{w}$, moreover $v \sim_4 w$ yields that $\hat{v} = \hat{w}$. Hence, $\hat{w}$ is a normal form of w.*

Finally, it is going to be shown that the length of this normal form is the least possible up to a constant factor. Let $f_k(n)$ be the number of $\sim_k$-equivalence classes over an n-letter alphabet. In [4] we proved that $\log f_k(n) = \Theta_k(n^{\frac{k+1}{2}})$, if k is odd and $\log f_k(n) = \Theta_k(n^{\frac{k}{2}} \log n)$, if k is even. From these estimates it follows immediately that there exists a word such that in its $\sim_k$-equivalence class the length of every word is at least $\Omega_k\left(\frac{n^{\frac{k+1}{2}}}{\log n}\right)$, if k is odd and at least $\Omega_k(n^{\frac{k}{2}})$, if k is even. Hence, there must be a word w such that in its $\sim_4$-equivalence class even the shortest word has length at least $\Omega(n^2)$. The normal form defined in this paper has length $O(n^2)$, therefore, up to a constant factor its length is the least possible.

References

[1] F. Blanchet-Sadri: *Games equations and dot-depth hierarchy*, Comput. Math. Appl. **18** (1989) 809–822.

[2] F. Blanchet-Sadri: *Equations and monoids varieties of dot-depth one and two*, Theoret. Comput. Sci. **123** (1994) 239–258.

[3] P. Karandikar, M. Kufleitner, P. Schnoebelen: *On the index of Simon's congruence for piecewise testability*, Information Processing Letters **15(4)** (2015) 515–519.

[4] K. Kátai-Urbán, P. P. Pach, G. Pluhár, A. Pongrácz, Cs. Szabó: *On the word problem for syntactic monoids of piecewise testable languages,* Semigroup Forum **84(2)** (2012) 323–332.

[5] J. E. Pin: *Varieties of Formal Languages*, North Oxford Academic, Plenum, 1986.

[6] I. Simon: *Piecewise testable events*, in Proc. 2nd GI Conf., Lect. Notes in Comput. Sci. **33** (1975) 214–222.

[7] J. Stern: *Complexity of some problems from the theory of automata*, Inform. and Control **66** (1985) 163–176.