NORMAL FORMS UNDER SIMON'S CONGRUENCE

Péter Pál Pach[1]

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 $\sim_k$. The length of this normal form is the shortest possible. Moreover, a canonical solution of the equation $pwq \sim_k r$ is also shown (the words p, q, r are parameters), which can be viewed as a generalization of giving a normal form for $\sim_k$. In this paper, there can be found an algorithm with which the canonical solution can be determined in $O((L+n)n^k)$ time, where L denotes the length of the word pqr and n is the size of the alphabet.

Key words and phrases: combinatorics of words, normal form, piecewise testable languages

1. Introduction

A large class of languages is the family of piecewise testable languages, which has been deeply studied in formal language theory, see for example, [8] or [**?**]. 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 scattered 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 $k = 3$. In [6] a normal form was given when $k = 4$. The new idea was to investigate a more general question, namely, to determine a canonical solution of the equation $pwq \sim_k r$. Here the words p, q, r are parameters and our aim is to solve the equation for the variable w (which is also a word). With the help of a canonical solution for the equations of the form $pwq \sim_2 r$ a normal form was shown for $\sim_4$. More generally, it has been proved 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 $k+2$. In this paper our goal is to define a canonical solution for

[1]Department of Computer Science and Information Theory, Budapest University of Technology and Economics, 1117 Budapest, Magyar tudósok körútja 2., Hungary
ppp@cs.bme.hu. This research was supported by the National Research, Development and Innovation Office NKFIH (Grant Nr. PD115978 and K124171) and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

the equations $pwq \sim_k r$ for arbitrary k. Furthermore, we are going to show that the canonical solution is a solution having minimal length. The special case, when $p = q = \textit{empty word}$ gives a normal form for the word r. Moreover, in the last section an algorithm is presented for finding the canonical solution. It is also shown that the canonical solution can be found in $O((L+n)n^k)$ deterministic time, where L is the length of pqr and n is the size of the alphabet.

2. Preliminaries

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

$$X^*x_1X^*x_2X^*\dots X^*x_lX^*, \ \textit{where } x_1,\dots,x_l \in X, \ 0 \leq l \leq k.$$

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

Simon [8] 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 $k > 3$. See Pin's textbook [7] for further details.

In this paper the alphabet X is going to be an n-element set (for some $n \in \mathbb{N}$). For a word w let $|w|$ denote its length. For a word w let us denote the set of its subwords of length at most k by $(w)_k$:

$$(w)_k = \{u : \ u \leq w \ \textit{and} \ |u| \leq k\}.$$

This way $w_1 \sim_k w_2$ holds 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 $c(w)$ for any $k \geq 1$. Let w' denote the word in which only the first and final occurences of the letters of w are kept, and the others are deleted. (Specially, if a letter occurs only once, we keep it.) Note that the word w' has length at most $2n$ and $(w)_2 = (w')_2$.

In Corollary 13, Proposition 14 and Section 4 we are going to use the O, Ω, Θ notions. For functions $f(k,n), g(k,n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ we write

- $f(k,n) = O_k\left(g(k,n)\right)$ if for every k there exists some constant d_k such that

$f(k,n) \leq d_k g(k,n)$ for every sufficiently large n (if d_k does not depend on k we simply write O),

- $f(k,n) = \Omega_k(g(k,n))$ if for every k there exists some constant $c_k > 0$ such that $f(k,n) \geq c_k g(k,n)$ for every sufficiently large n (if c_k does not depend on k we simply write Ω),
- $f(k,n) = \Theta_k(g(k,n))$ if for every k there exists some constants $0 < c_k, d_k$ such that $c_k g(k,n) \leq f(k,n) \leq d_k g(k,n)$ for every sufficiently large n (if c_k and d_k do not depend on k we simply write Θ).

3. Solving the equation $pwq \sim_k r$

In this section our aim is to give a well-defined (*canonical*) solution of the equation $pwq \sim_k r$, or equivalently $(pwq)_k = (r)_k$ if a solution exists. Here the words p, q, r are parameters, and we would like to solve the equation for w. The solution should only depend on the equivalence classes of the words p, q, r, that is, on $(p)_k$, $(q)_k$ and $(r)_k$. The canonical solution that we define is going to be denoted by $\overline{w} = \overline{w}^{(p,q,r)}$. This approach is a generalization of finding a normal form for the words under $\sim_k$, since if p and q are the empty word, then the canonical solution of the equation $w \sim_k r$ is a normal form for the word r. Furthermore, this problem can be viewed as finding a canonical form for the word r in such a way that it begins with p and ends with q.

At first assume that $k = 1$. The equation $pwq \sim_1 r$ holds if and only if $c(pwq) = c(r)$, hence w is a solution if and only if $c(r) \setminus (c(p) \cup c(r)) \subseteq c(w) \subseteq c(r)$. Let $\overline{w} = \overline{w}^{(p,q,r)}$ be the word obtained by writing the elements of $c(r) \setminus (c(p) \cup c(q)) \subseteq c(w)$ in alphabetical order after each other. Then $\overline{w}$ solves the equation and its length is the shortest possible.

From now, assume that $k \geq 2$. We are looking for a word w for which $(pwq)_k = (r)_k$. At first we are going to find a possible choice for the word $(pwq)' = y = y_1 y_2 \dots y_t$. (Here $t \leq 2n$.) Note that, even if $pwq \sim_k r$, it is possible that $(pwq)' \neq r'$ and it is also possible that there are more than one possibilities for $y = (pwq)'$. (Since different solutions in w can result in different y's.)

Let us write pwq as $pwq = y_1 u_1 y_2 u_2 \dots y_t$. After finding an appropriate y, the words $u_1, u_2, \dots, u_{t-1}$ are going to be defined in such a way that the word $y_1 u_1 y_2 u_2 \dots y_t$ begins with p, ends with q (these parts do not overlap) and in the middle the solution w could be found.

It can easily be seen that y and the $\sim_{k-2}$ equivalence classes of the words $u_1, u_2, \dots, u_{t-1}$ together determine $(pwq)_k$.

Proposition 1. *If $(y_1 u_1 y_2 u_2 \dots y_t)' = y_1 y_2 \dots y_t = y$, then $(y_1 u_1 y_2 \dots y_t)_k$ is determined by y and $(u_1)_{k-2}, \dots, (u_t)_{k-2}$. (Note that here y contains the first and last*

(or unique) appearances of the letters. So each letter can appear at most twice in y. If a letter appears just once, then it can't appear elsewhere. If a letter appears twice in y, say at positions y_i, y_j (with $i < j$), then it can appear in $u_i, u_{i+1}, \ldots, u_{j-1}$, but can't appear anywhere else.)

We give a proof of this statement in the spirit of [5].

Proof. Let us suppose that $z = z_1 z_2 z_3$ 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_α be the first appearance of the letter z_1 in y and y_β the last appearance of the letter z_3 in y. Clearly, $z \in (y_1 u_2 y_2 u_2 \ldots y_t)_k$ if and only if $\alpha < \beta$ and $z_2 \in (u_\alpha y_{\alpha+1} \ldots u_{\beta-1})_{k-2}$. Therefore, the set of the at most k-letter subwords of $y_1 u_1 y_2 u_2 \ldots y_t$ is determined by y and $(u_1)_{k-2}, \ldots, (u_t)_{k-2}$. □

Now we concentrate on $y = (pwq)'$. The word y is going to be defined with the help of the equation $pwq \sim_k r$. At first we are going to characterize the possible s' words if $s \sim_k r$.

For $a \in c(r)$ let

$$R(a) = R_r^{k-1}(a) = a^{-1}(r)_k = \{w \,|\, aw \leq r \text{ and } |w| \leq k-1\}$$

and

$$L(a) = L_r^{k-1}(a) = (r)_k a^{-1} = \{w \,|\, wa \leq r \text{ and } |w| \leq k-1\}.$$

The sets $R_r^{k-1}(a)$ and $L_r^{k-1}(a)$ are determined by $(r)_k$. Since, the set $R_r^{k-1}(a)$ can be obtained by taking all the words in $(r)_k$ starting with the letter a and deleting their first letter, namely, a. Similarly, $L_r^{k-1}(a)$ can be obtained by taking all the words in $(r)_k$ ending with letter a and deleting their last letter, namely, a.

A letter a occurs exactly once in r, or equivalently in r', if and only if aa is not a subword of r. Let $a \approx b$, if $R_r^{k-1}(a) = R_r^{k-1}(b)$ for $a, b \in c(r)$ which occur at least twice in w. Clearly, $\approx$ is an equivalence relation on a subset of the set $c(r)$. For a letter $a \in c(r)$ which occurs at least twice in r the $\approx$ class of a is called the R-block of a. The L-blocks are defined dually. If $a \in c(r)$ occurs exactly once in r, then $\{a\}$ is called a U-block. The R-blocks, the L-blocks and the U-blocks are called blocks. The set of the blocks of r' depends only on $(r)_k$. Moreover, it is going to be shown that the elements of each block are consecutive letters in r' and the order of the blocks in r' is determined by $(r)_k$, as well. Therefore, the word r' is determined by $(r)_k$ up to the order of the letters within the R-blocks and the L-blocks. Accordingly, for each $1 \leq i \leq t$ the block of the letter y_i is determined.

Proposition 2. *The first (last) appearences of the letters of an R-block (or L-block) are consecutive letters of r'. The order of the blocks in r' is uniquely determined by $(r)_k$.*

Proof. This proposition is a straightforward consequence of the following observations:

(i) For $a, b \in c(r)$ let y_i be the first appearance of the letter a in r' and y_j the last appearance of the letter b in r'. Then $i < j$ if and only if $ab \in (r)_k$,

(ii) For $a, b \in c(r)$ let y_i be the first appearance of the letter a in r' and y_j the first appearance of the letter b in r'. If $R_r^{k-1}(b) \subsetneq R_r^{k-1}(a)$, then $i < j$.

(iii) For $a, b \in c(r)$ let y_i be the last appearance of the letter a in r' and y_j the last appearance of the letter b in r'. If $L_r^{k-1}(a) \subsetneq L_r^{k-1}(b)$, then $i < j$.

□

Therefore, the length of the word $(pwq)' = y = y_1 y_2 \dots y_t$, the block of each y_i and its type (R-, L- or U-block) are determined by $(r)_k$. Moreover, each block contains consecutive letters of y, so up to the orders of the letters within the blocks the word y is determined. Now, our aim is to find which part of $y = (pwq)'$ belongs to p, w, q, respectively. In order to do this the content of w is investigated.

Clearly, $c(w) \subseteq c(r)$. Let C contain those letters a for which there exists a subword $v_1 a v_2$ of r of length at most k such that $v_1 a \not\leq p$ and $a v_2 \not\leq q$. We claim that for any solution w we have $C \subseteq c(w)$. Furthermore, by keeping only such letters of a solution w that are in C we still obtain a solution.

Proposition 3. *Let*

$$C = \{a \mid \exists v_1, v_2 : v_1 a v_2 \leq r, |v_1 a v_2| \leq k, v_1 a \not\leq p, a v_2 \not\leq q\}.$$

If $pwq \sim_k r$, then $C \subseteq c(w)$. Moreover, if the word w_C is obtained from w by keeping only the letters of C, then $pw_Cq \sim_k r$.

Proof. Let us assume that $a \in C$. Then there exist words v_1, v_2 such that $v_1 a v_2 \in (r)_k = (pwq)_k$ and $v_1 a \not\leq p$, $a v_2 \not\leq q$. Let us choose a certain occurence of $v_1 a v_2$ in the word pwq. The letter a in the middle of $v_1 a v_2$ can not belong to p, since it would mean that $v_1 a \leq p$. Similarly, the letter a can not belong to q, neither. Therefore, $a \in c(w)$, thus $C \subseteq c(w)$ is proved.

To prove that w_C solves the equation it is enough to show that if $pwq \sim_k r$ and $a \in c(w) \setminus C$, then by deleting one appearance of a from w we still obtain a word w^* satisfying $pw^*q \sim_k r$. By this, the appearances of the letters in $c(r) \setminus C$ can be deleted from w one by one. Clearly, $(pw^*q)_k \subseteq (pwq)_k = (r)_k$, so it suffices to prove that all subwords of $(r)_k = (pwq)_k$ are subwords of pw^*q, as well. So let us assume that $v \leq r$ for a word v of length at most k. Only a single letter (a) was deleted from w, hence v can be written as $v = v_1 a v_2$ and w can be written as $w = w_1 a w_2$ such that $v_1 \leq p w_1$ and $v_2 \leq w_2 q$. As $a \notin C$ either $v_1 a \leq p$ or $a v_2 \leq q$. We may assume that $v_1 a \leq p$, the other case can be handled similarly. Now, $v_1 a \leq p$ and $v_2 \leq w_2 q$, therefore $v = v_1 a v_2 \leq p w_1 w_2 q = pw^*q$. □

It is obtained that each element of C must appear in each solution w of the equation $pwq \sim_k r$. Moreover, any letter not contained in C can be deleted from any solution w, and we still obtain a solution. As a consequence, from now on, it is assumed that $c(w) = C$. If $pwq \sim_k r$ is solvable (in w), then there is a solution for which $c(w) = C$ holds, as well.

Now, we are able to decide which part of $(pwq)' = y_1 y_2 \dots y_t$ belongs to p, w, q, respectively. Firstly, if $a = y_i$ is a unique appearance in r, then we can mark it in p or q (if $a \in c(p)$ or $a \in c(q)$), otherwise (that is, if $a \notin c(p) \cup c(q)$) it belongs to $c(w)$. As the order of the blocks is determined, the index i is also determined by the letter a. Secondly, suppose that $a = y_i$ is a first appearance. If $a \in c(p)$, then we can mark it in p: It is the first appearance of the letter a in p. If $a \notin c(p)$, but $a \in C$, then the first appearance of a belongs to w. Finally, if $a \in c(r) \setminus (c(p) \cup C)$, then the first appearance of a belongs to q, and we can find it in q: It is the first appearance of the letter a in q. When $a = y_i$ is a final appearance, then the place of y_i can be found in pwq dually to the previously observed case.

Therefore, for each letter it is determined which part of pwq contains the first and final appearance of it. As p and q are given words, these appearances can be marked in them, and consequently the order of the letters of $y = (pwq)'$ is determined in these two parts. We also know which part of y belongs to w, but here – up to now – only the order of the blocks is known. Let $y_1 y_2 \dots y_i$ be the part contained in p, $y_{i+1} y_{i+2} \dots y_j$ be the part contained in w and $y_{j+1} y_{j+2} \dots y_t$ be the part contained in q. The letters of $(pwq)' = y$ divide the words p and q into several parts:

$$p = y_1 u_1 y_2 \dots y_i u_{i,1}$$

$$q = u_{j,2} y_{j+1} u_{j+1} y_{j+2} \dots y_t$$

Here, all words are determined by now, and we are looking for w in the following form:

$$w = u_{i,2} y_{i+1} u_{i+1} \dots y_j u_{j,1}.$$

With these notions, $pwq = y_1 u_1 y_2 u_2 \dots y_t$, where $u_i = u_{i,1} u_{i,2}$ and $u_j = u_{j,1} u_{j,2}$. However, in w only the blocks of $y_{i+1}, \dots, y_j$ are determined yet. At first an appropriate choice for the order of these letters is going to be given, and after that the gaps between them are going to be filled in, appropriate u_l words are going to be found.

Let us summarize the steps that we have done until now:

Step 1. Find the blocks of $(r)_k$. Determine the order of the blocks in y. Determine C and which part of y is contained in p, w, q, respectively.

Now we continue with finding a good ordering of the letters for each block. Let $B = \{b_1, b_2, \dots, b_v\}$ be a block. If it is a U-block, then its size is 1 and there is only one ordering. Let us assume that B is an R-block. (The dual case when B

is an L-block can be handled similarly.) In this case $R_r^{k-1}(b_1) = R_r^{k-1}(b_2) = \cdots = R_r^{k-1}(b_v) =: R$. If the block B has no elements in w, then the order of the elements of this block is determined by p and q. So it can be assumed that at least one element of B is contained in w. There are four cases:

(i) The whole block B belongs to w.

(ii) Some part of B belongs to p and some part of it to w.

(iii) Some part of B belongs to w and some part of it to q.

(iv) B has elements in p, w and q, as well.

We deal with these four cases simultaneously. The letters of B are consecutive letters in $y = y_1 y_2 \dots y_t$. Let us assume that for a solution w the order of the elements of B in the R-block contained in y is $b_1, b_2, \dots, b_v$. Note that this order is not necessarily uniquely determiend by $(p)_k, (q)_k$ and $(r)_k$. Our aim is to find an appropriate ordering in a canonical way.

Let us write pwq as

$$pwq = z_0 b_1 z_1 b_2 z_2 \dots z_{v-1} b_v z_v,$$

where the indicated appearances of $b_1, b_2, \dots, b_v$ are all first (that is, left-most) appearances. Let us assume that $b_1, b_2, \dots, b_{\alpha-1}$ belong to p and $b_\alpha, b_{\alpha+1}, \dots, b_\beta$ belong to w, finally $b_{\beta+1}, \dots, b_v$ belong to q. In cases (i) and (iii), we have $\alpha = 1$; and in cases (i) and (ii), $\beta = v$ holds. Here $p = z_0 b_1 z_1 \dots b_{\alpha-1} z_{\alpha-1,1}$, $w = z_{\alpha-1,2} b_\alpha z_\alpha \dots b_\beta z_{\beta,1}$ and $q = z_{\beta,2} b_{\beta+1} z_{\beta+1} \dots b_v z_v$, where $z_{\alpha-1,1} z_{\alpha-1,2} = z_{\alpha-1}$ and $z_{\beta,1} z_{\beta,2} = z_\beta$.

As the next step we prove that if w is a solution, then by replacing $z_{\alpha-1,2}, z_\alpha, z_{\alpha+1}, \dots, z_{\beta-1}, z_{\beta,1}$ by the empty word, the obtained word w^B still satisfies $pw^Bq \sim_k r$. Let us use the notions $z_l^B = z_l$ if $l \in \{1, 2, \dots, \alpha-2, \beta+1, \beta+2, \dots, v-1\}$ and $z_\alpha^B = z_{\alpha+1}^B = \cdots = z_{\beta-1}^B =$ *empty word* and $z_{\alpha-1}^B = z_{\alpha-1,1}$, $z_\beta^B = z_{\beta,2}$.

Proposition 4. *Let us assume that B is an R-block and $pwq \sim_k r$. Let w^B be the word obtained from w in such a way that every letter between two first appearances from the block B belonging to w, are deleted. Then $pw^Bq \sim_k r$.*

Proof. We use the previously introduced notions. Clearly, pw^Bq can be written as $pw^Bq = z_0 b_1 z_1^B \dots z_{v-1}^B b_v z_v$, where for each $1 \le l \le v-1$ we have $z_l^B \le z_l$. As pw^Bq is obtained from pwq by deleting some letters, $(pw^Bq)_k \subseteq (pwq)_k$. Now we show that $(pw^Bq)_k \supseteq (pwq)_k$ also holds. Let $u \in (pwq)_k$ and let us write u as $u = u_1 u_2$, where u_1 is the first letter of u, and consequently the length of u_2 is at most $k-1$. The word u might appear more than once in pwq, let us choose a certain occurence of it. It can be supposed that the first letter of u, that is, u_1 is a first appearance in pwq. If this preceeds the first appearances of the letters from B, then in pwq

the part $z_1b_2\dots z_{v-1}b_vz_v$ contains at most $k-1$ letters of u. B is an R-block, so $(z_1b_2\dots z_{v-1}b_vz_v)_{k-1}=R(b_1)=R(b_v)=(z_v)_{k-1}$, so this part is contained in z_v in pw^Bq, as well. Since z_0b_1 has not changed, we obtained that $u\in(pw^Bq)_k$ holds. If u_1 is in B, then $u_2\in(z_1b_2\dots b_vz_v)_{k-1}=R(b_1)=R(b_v)=(z_v)_{k-1}$. Since z_v is at the end of pw^Bq, again we have $u\in(pw^Bq)_k$. Finally, if u_1 is a first appearance after the R-block B, then $u\le z_v$, so $u\in(pw^Bq)_k$. Therefore, we proved that $pw^Bq\sim_k pwq$. □

So we can suppose that those elements of B that belong to w are consecutive letters in w. Now, let us determine which permutations of $b_\alpha,b_{\alpha+1},\dots,b_\beta$ give a solution, that is, which choice of

$$pw^\pi q=z_0b_1z_1^Bb_2z_2^B\dots b_{\alpha-1}z_{\alpha-1}^Bb_{\pi(\alpha)}b_{\pi(\alpha+1)}\dots b_{\pi(\beta)}z_\beta^Bb_{\beta+1}\dots z_{v-1}^Bb_vz_v$$

we have

$$pw^\pi q\sim_k\sim_k pw^Bq(\sim_k r).$$

For the simplicity of the notions let us extend π to a permutation of the numbers $\{1,2,\dots,v\}$ in such a way that $1,2,\dots,\alpha-1,\beta+1,\dots,v$ are all fixed points of the extended permutation, which is also denoted by π.

Let us call a permutation π *good* if the rearranged word w^π still solves the equation. We are going to show that there are two cases: either every π is good, or π is good if and only if $b_1=b_{\pi(1)}$.

Lemma 5. *We claim that a permutation π is good if and only if*

$$(z_1^Bb_{\pi(2)}z_2^B\dots b_{\pi(\alpha-1)}z_{\alpha-1}^Bb_{\pi(\alpha)}b_{\pi(\alpha+1)}\dots b_{\pi(\beta)}z_\beta^Bb_{\pi(\beta+1)}\dots z_{v-1}^Bb_{\pi(v)}z_v)_{k-1}=\\=(z_v)_{k-1}.$$

Proof. If π is good, then $R^{k-1}_{pw^Bq}(b_{\pi(1)})=R^{k-1}_{pw^Bq}(b_{\pi(v)})$, so the condition is necessary. To prove that it is sufficient we have to show that

$$(z_0b_1z_1^Bb_2\dots z_{v-1}^Bb_vz_v)_k=(z_0b_{\pi(1)}z_1^Bb_{\pi(2)}\dots z_{v-1}^Bb_{\pi(v)}z_v)_k.$$

Clearly, it is enough to prove that

$$(b_1z_1^Bb_2\dots z_{v-1}^Bb_vz_v)_k=(b_{\pi(1)}z_1^Bb_{\pi(2)}\dots z_{v-1}^Bb_{\pi(v)}z_v)_k.$$

Let u be a word of length at most k. Let $u=u_1u_2$, where u_1 is the first letter of u, and consequently the length of the word u_2 is at most $k-1$. At first assume that $u_1\in c(z_1^Bz_2^B\dots z_{v-1}^B)\cup B$. If $u\le b_1z_1^Bb_2\dots z_{v-1}^Bb_vz_v$, then $u_2\in(z_1^Bb_2\dots z_{v-1}^Bb_vz_v)_{k-1}=R(b_1)=(z_v)_{k-1}$. If $u\le b_{\pi(1)}z_1^Bb_{\pi(2)}\dots z_{v-1}^Bb_{\pi(v)}z_v$, then $u_2\in(z_1^Bb_{\pi(2)}z_2^B\dots z_{v-1}^Bb_{\pi(v)}z_v)_{k-1}=(z_v)_{k-1}$. Conversely, if $u_2\in(z_v)_{k-1}$, then clearly $u\le b_1z_1^Bb_2\dots z_{v-1}^Bb_vz_v$ and $u\le b_{\pi(1)}z_1^Bb_{\pi(2)}\dots z_{v-1}^Bb_{\pi(v)}z_v$. Therefore,

$$u=u_1u_2\le b_1z_1^Bb_2\dots z_{v-1}^Bb_vz_v\iff u_2\le z_v\iff\\\iff u=u_1u_2\le b_{\pi(1)}z_1^Bb_{\pi(2)}\dots z_{v-1}^Bb_{\pi(v)}z_v,$$

so in this case the statement is true. Finally, if $u_1 \notin c(z_1^B z_2^B \dots z_{v-1}^B) \cup B$, then clearly

$$u \leq b_1 z_1^B b_2 \dots z_{v-1}^B b_v z_v \iff u \leq z_v \iff u \leq b_{\pi(1)} z_1^B b_{\pi(2)} \dots z_{v-1}^B b_{\pi(v)} z_v.$$

So, it has been shown that π is good if and only if

$$(z_1^B b_{\pi(2)} z_2^B \dots z_{v-1}^B b_{\pi(v)} z_v)_{k-1} = (z_v)_{k-1}.$$

□

Now we prove that only $\pi(1)$ determines whether π is good, or not:

Lemma 6. *If $\pi(1) = 1$, then π is good.*

Proof. Let us suppose that

$$u = u_1 u_2 \dots u_{k-1} \leq z_1^B b_{\pi(2)} z_2^B \dots z_{v-1}^B b_{\pi(v)} z_v.$$

It can be supposed that u_1 is a first appearance (in this word) and u_{k-1} is a final (or unique) appearance. As B is an R-block, $B \subseteq c(z_v)$. Moreover, since the letters of B are consecutive in $y = (pwq)'$, there are no final (or unique) appearances in $z_1^B, z_2^B, \dots, z_{v-1}^B$, that is, all of the letters in them appear in z_v, as well. Hence, $u_{k-1} \leq z_v$. Let i be minimal such that $u_i \dots u_{k-1} \leq z_v$. If $i > 1$, then $u_{i-1} \in c(z_1^B b_{\pi(2)} \dots z_{v-1}^B b_{\pi(v)}) \subseteq c(z_1^B b_2 \dots z_{v-1}^B b_v)$, so $u_{i-1} u_i \dots u_{k-1} \in R(b_1) = R(b_v) = (z_v)_{k-1}$, which contradicts the minimality of i. Hence, $i = 1$ and $u \leq z_v$ holds. Therefore, Lemma 5 implies that π is good. □

We obtained that if $b_1, b_2, \dots, b_v$ is a good ordering of the elements of the block B, then all of the orderings where $\pi(1) = 1$ are also good. With other words, only the first letter of the block plays a role. If b_1 belongs to p, all choices for π are appropriate, since if the equation $pwq \sim_k r$ is solvable, then there exists at least one good choice. Therefore, for instance, π can be the lexicographical order on $\{\alpha, \alpha + 1, \dots, \beta\}$. From now on assume that B has no elements in p, that is, we deal with the case (i) or (iii).

Lemma 7. *If not all permutations π are good, then there is only one choice for $b_{\pi(1)}$ with which π is good. This unique good choice for $b_{\pi(1)}$ is the letter from B which first appearance is the latest in the $\sim_{k-1}$ normal form of z_v.*

Proof. Since the equation $pwq \sim_k r$ is solvable, there exists at least one good choice for the first element of the block B: b_1. Now suppose that π is not good, that is, the value of $\pi(1)$ is not appropriate. According to Lemma 5 there exists a word $u = u_1 u_2 \dots u_{k-1} \leq z_1^B b_{\pi(2)} z_2^B \dots z_{v-1}^B b_{\pi(v)} z_v$ for which $u \not\leq z_v$. Let i be minimal such that $u_i \dots u_{k-1} \leq z_v$. If $u_{i-1} \in c(z_1^B b_2 \dots z_{v-1}^B b_v)$, then $u_{i-1} u_i \dots u_{k-1} \in$

$R(b_1) = R = (z_v)_{k-1}$, which is a contradiction. Therefore, $u_{i-1} = b_1$. If there would be another good starting element in B (other than b_1), then with a similar reasoning it would yield that u_{i-1} would have to be that element, which is also a contradiction. Hence, if not all starting elements are good, then there is only one good starting element, which is now b_1. Our aim is to find a good ordering in the case when there is only one appropriate starting letter. Using the previous notions, it follows that $b_1 u_i \dots u_{k-1} \not\leq z_v$, but clearly $b_2 u_i \dots u_{k-1}, b_3 u_i \dots u_{k-1}, \dots, b_v u_i \dots u_{k-1} \in R = (z_v)_{k-1}$, therefore in z_v the first appearance of the letter b_1 is after the first appearances of $b_2, b_3, \dots, b_v$. Hence, the unique appropriate choice b_l for the starting element $b_{\pi(1)}$ can be obtained in the following way:

- Take the $\sim_{k-1}$ normal form of z_v.
- Find the first appearances of the letters in B within this normal form.
- Then b_l is the one which has the latest (that is, the right-most) first appearance.

□

Hence, we can set $b_{\pi(1)} := b_l$ if b_l does not belong to q. Note that $(z_v)_{k-1} = R$ is determined by p, q, r, so we managed to find a canonical starting letter of the block B. The order of the other elements of B is arbitrary, for instance, take the lexicographical order. If b_l belongs to q, then $b_{\pi(1)} \neq b_l$, meaning that all permutations π are good, so we can take the lexicographical order on $\{\alpha, \alpha+1, \dots, \beta\}$.

Hereafter, for finding a canonical solution for the equation $pwq \sim_k r$ the next step is the following:

Step 2. For every block B determine a good ordering π of the elements of B in $y = (pwq)'$.

Example 8. *We illustrate with an example that it can indeed happen that within a block B there is just one good choice for $b_{\pi(1)}$. Let $k = 3$ and $w = abcbcbca$, then a, b, c are within the same R-block, since*

$$R(a) = R(b) = R(c) = \{\textit{empty word}, a, b, c, ba, bb, bc, ca, cb, cc\}.$$

Lemma 7 claims that to determine the starting element of this R-block we have to:

- *take the $\sim_2$ normal form of bcbca, which is bcbca,*
- *find the first appearances of a, b, c:* **bc**bc**a** *(marked bold),*
- *then the unique appropriate starting letter is the one which has the latest (that is, the right-most) first appearance, that is: a.*

If one would try to find another word in the $\sim_k$ equivalence class of w in which the starting letter of this R-block is b (or c), then it would imply that $aa \in R(b)$ (or $aa \in R(c)$), leading to a contradiction.

Therefore, a well-defined $y = (pwq)'$ is obtained and in

$$w = u_{i,2}y_{i+1}u_{i+1}\dots y_j u_{j,1}$$

all the letters $y_{i+1}, y_{i+2}, \dots, y_j$ are chosen in an appropriate way. It only remains to find appropriate $u_{i,2}, u_{i+1}, \dots, u_{j,1}$ subwords. Note that for $k = 2$ this last part is not needed, since $pwq \sim_2 (pwq)'$.

Now we know that $y = y_1y_2\dots y_t$ satisfies that w can be chosen in such a way that

$$pwq = y_1u_1y_2u_2\dots y_t \sim_k r,$$

and here in $p = y_1u_1y_2\dots y_iu_{i,1}$ and $q = u_{j,2}y_{j+1}u_{j+1}y_{j+2}\dots y_t$ every u_l subword is known, so the words that should be defined are all in

$$w = u_{i,2}y_{i+1}u_{i+1}\dots y_j u_{j,1},$$

namely, $u_{i,2}, u_{i+1}, \dots u_{j-1}, u_{j,1}$. (Clearly, $u_i = u_{i,1}u_{i,2}$ and $u_j = u_{j,1}u_{j,2}$.)

According to Proposition 1 the words $y = (pwq)'$ and $(u_1)_{k-2}, \dots, (u_{t-1})_{k-2}$ together determine $(pwq)_k$, 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 (where $a < b$) the following holds (note that we know that an appropriate choice exists):

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

Let r_0 be the subword of r containing every letter of r between the first appearance of y_a and the last appearance of y_b (the first y_a and the last y_b is not included). Then $(r_0)_{k-2} = M_{y_a,y_b}(r)$.

At first we determine an order in which the words $(u_l)_{k-2}$ are going to be defined. For $1 \le l \le t-1$ let n_l be the total number of first appearances in $\{y_{l+1}, \dots, y_t\}$ and last appearances in $\{y_1, \dots, y_l\}$. We define the u_l words in increasing order according to n_l. Suppose that for some l the words u_m for which $n_m < n_l$ are already defined. We show that now u_l is definable, as well. Let $\alpha \le l$ be maximal such that y_α is a first appearance and $l+1 \le \beta$ be minimal such that y_β is a last appearance. Since $y_{\alpha+1}, y_{\alpha+2}, \dots, y_l$ are all last appearances and $y_{l+1}, y_{l+2}, \dots, y_{\beta-1}$ are all first appearances, $\max(n_\alpha, n_{\alpha+1}, \dots, n_{l-1}, n_{l+1}, n_{l+2}, \dots, n_{\beta-1}) < n_l$, so $u_\alpha, u_{\alpha+1}, \dots, u_{l-1}, u_{l+1}, u_{l+2}, \dots, u_{\beta-1}$ are already defined (in an appropriate way). Let $p_0 = u_\alpha y_{\alpha+1}u_{\alpha+1}\dots y_l$, $q_0 = y_{l+1}u_{l+1}\dots u_{\beta-1}$ and $(r_0)_{k-2} = M_{y_\alpha,y_\beta} = \{m \mid y_\alpha m y_\beta \in (r)_k\}$. The word u_l has to satisfy the equation $(p_0u_lq_0)_{k-2} = (r_0)_{k-2}$, so let us choose u_l as the canonically defined solution of this equation: $u_l := \overline{u}_l = \overline{u}_l^{(p_0,q_0,r_0)}$. Now, we show that for any appropriate choice of the words

u_m, that is, for any choice for which all the equations of the form (1) hold if we replace u_l by the previously defined $\overline{u}_l$, they will still hold. It means that by setting u_l to be $\overline{u}_l$ we can not make a "mistake".

When we check the equation $M_{y_a,y_b} = (u_a y_{a+1} \dots u_{b-1})_{k-2}$ for some first appearance y_a and last appearance y_b (satisfying $a < b$), then the choice of u_l only plays a role if $a \leq l < b$. This yields $a \leq \alpha$ and $\beta \leq b$. In the special case when $a = \alpha$ and $l = \beta$, according to the definition of $\overline{u_l}$, we have $(u_\alpha y_{\alpha+1} \dots u_{\beta-1})_{k-2} = M_{y_\alpha,y_\beta}$. Here, M_{y_α,y_β} is determined by $(pwq)_k = (r)_k$, therefore $(u_\alpha y_{\alpha+1} \dots u_{\beta-1})_{k-2}$ is also determined by $(r)_k$. Using this observation we obtain that for arbitrary $a \leq \alpha$ and $\beta \leq b$ the right hand side of

$$(u_a y_{a+1} \dots u_{b-1})_{k-2} = (u_a y_{a+1} \dots y_\alpha)_{k-2} (u_\alpha y_{\alpha+1} \dots u_{\beta-1})_{k-2} (y_\beta u_\beta \dots u_{b-1})_{k-2},$$

does not depend on the choice of u_l, the only restriction for u_l is that it has to satisfy $(u_\alpha y_{\alpha+1} \dots u_{\beta-1})_{k-2} = M_{y_\alpha,y_\beta}$. Hence, we can set $u_l := \overline{u_l}$.

There are two special cases: $l = i$ and $l = j$. When $l = i$, then a slight modification is needed in the definition of p_0: In this case $p_0 = u_\alpha y_{\alpha+1} u_{\alpha+1} \dots y_i u_{i,1}$, that is, a word $u_{i,1}$ (the beginning of u_i) has to be written at the end of p_0, since $u_{i,1}$ is determined by p. Similarly, when $l = j$, the definition of q_0 should be modified in the following way: $q_0 = u_{j,2} y_{j+1} u_{j+1} \dots u_{\beta-1}$. With these modifications the above arguments are valid in these two special cases, as well.

Therefore, one by one the words u_l can be defined with the help of a canonical form of a solution of the equations of the form $(p_0 u q_0)_{k-2} = (r_0)_{k-2}$, and finally the normal form $\overline{w} = y_1 \overline{u}_1 y_2 \overline{u}_2 \dots \overline{u}_{t-1} y_t$ is obtained.

Thus we arrived at the final step:

Step 3. Find appropriate u_l words.

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

Proposition 9. *Let $k \in \mathbb{N}$. Let p, q, r, p', q', r' be words and suppose that the equation $pwq \sim_k r$ has a solution. Then $p\overline{w}^{(p,q,r)} q \sim_k r$. If $p \sim_k p', q \sim_k q', r \sim_k r'$, then $\overline{w}^{(p,q,r)} = \overline{w}^{(p',q',r')}$. Hence, $\overline{w}^{(p,q,r)}$ is a canonical form of a solution for the equation $pwq \sim_k r$.*

Note that $w \not\sim_k \overline{w}^{(p,q,r)}$ is possible even if $pwq \sim_k r$; and that the latter condition implies $pwq \sim_k p\overline{w}^{(p,q,r)} q$.

As a corollary a normal form is obtained for $\sim_k$. Let $\hat{r}$ be the canonical solution of the equation $w \sim_k r$, that is, when p and q are the empty word. Then $\hat{r}$ is a normal form for r.

Corollary 10. *Let $k \in \mathbb{N}$. Let r and s be two words. Then $r \sim_k \hat{r}$, moreover $r \sim_k s$ yields that $\hat{r} = \hat{s}$. Hence, $\hat{r}$ is a normal form of r.*

Finally, it is going to be shown that the length of this normal form is the least possible.

Theorem 11. *Let $k \in \mathbb{N}$. The length of the canonical solution $\overline{w} = \overline{w}^{p,q,r}$ of the equation $pwq \sim_k r$ is minimal.*

Proof. We prove the statement by induction on k. For $k = 1$ the length of $\overline{w} = \overline{w}^{p,q,r}$ is the cardinality of the set $c(r) \setminus (c(p) \cup c(q))$. Moreover, a word w is a solution of the equation $pwq \sim_1 r$ if and only if $c(r) \setminus (c(p) \cup c(q)) \subseteq c(w) \subseteq c(r)$. Therefore, the length of $\overline{w}$ is minimal.

For $k = 2$ let w^* be a solution of $pwq \sim_2 r$ having minimal length. According to Proposition 3, $C \subseteq c(w^*)$, furthermore w^*_C is also a solution, so by the minimality of $|w^*|$ and $w^*_C \leq w^*$ it follows that $c(w^*) = C = c(\overline{w})$. It has been also shown that if w is a solution of $pwq \sim_2 r$, then the content of w determines which part of $(pwq)'$ belongs to p, w, q, respectively. Hence, if from the word $y = (p\overline{w}q)'$ exactly $y_{i+1}y_{i+2}\dots y_j$ is the part which is contained in $\overline{w}$, then for the word $y^* = (pw^*q)'$ it also holds that exactly the part $y^*_{i+1}y^*_{i+2}\dots y^*_j$ is contained in w^*. The length of $\overline{w}$ is $j - i$ and the length of w^* is at least $j - i$, so $\overline{w}$ has minimal length.

Now assume that $k \geq 3$ and that the statement is proved up to $k - 1$. Let w^* be a solution of $pwq \sim_k r$ of minimal length. As $pw^*q \sim_k p\overline{w}q \sim_k r$, the length of the words $y = (p\overline{w}q)'$ and $y^* = pw^*q$ is the same, moreover the multiset of the letters of y and y^*, the blocks and the order of the blocks is the same, as well. According to Proposition 3, $c(\overline{w}) = C \subseteq c(w^*)$, moreover $pw^*_Cq \sim_k r$ also holds for the word $w^*_C \leq w^*$, so the minimality of $|w^*|$ implies that $w^*_C = w^*$, that is, $c(w^*) = C = c(\overline{w})$. Since $c(w^*)$ determines which part of $(pw^*q)' = y^*$ belongs to p, w^*, q respectively, the multiset of the letters of y contained in $\overline{w}$ and the letters of y^* contained in w^* is the same. Therefore, $\overline{w}$ and w^* can be written as

$$\overline{w} = u_{i,2}y_{i+1}u_{i+1}\dots y_ju_{j,1},$$

$$w^* = u^*_{i,2}y^*_{i+1}u^*_{i+1}\dots y^*_ju^*_{j,1}.$$

Let $i + 1 \leq l \leq j - 1$ be arbitrary, it is going to be shown that $|u^*_l| \geq |u_l|$. Let us recall that u_l satisfies the equation $p_0u_lq_0 \sim_{k-2} r_0$, where

$$p_0 = u_\alpha y_{\alpha+1}u_{\alpha+1}\dots y_l,$$

$$q_0 = y_{l+1}u_{l+1}\dots u_{\beta-1}$$

and

$$(r_0)_{k-2} = M_{y_\alpha, y_\beta} = \{m \mid y_\alpha m y_\beta \in (r)_k\}.$$

If we manage to show that $(p_0)_{k-2}, (q_0)_{k-2}, (r_0)_{k-2}$ are determined by p, q, r, then by the induction hypothesis applied for $k - 2$ it follows that $|u_l| \leq |u^*_l|$, since $p_0 \sim_{k-2} p^*_0, q_0 \sim_{k-2} q^*_0, r_0 \sim_{k-2} r^*_0$. At first we show that if $a = y_\gamma$ is a first (or unique) appearance and $b = y_\delta$ is a final (or unique) appearance, then $M_{ab} =$

$M_{y_\gamma y_\delta} = \{m \mid amb \in (r)_k\}$ is determined by the R-block (or U-block) of $y_\gamma = a$ and the L-block (or U-block) of $y_\delta = b$. From the equations

$$\{m \mid mb \in R_r^{k-1}(a)\} = M_{ab} = \{m \mid am \in L_r^{k-1}(b)\},$$

it follows that M_{ab} is determined by b and the R-block of a, moreover, by a and the L-block of b, as well. Hence, M_{ab} depends only on the R-block of a and the L-block of b. Therefore, $(r_0)_{k-2}$ is determined, furthermore, $M_{y_\alpha y_l} = (u_\alpha y_{\alpha+1} u_{\alpha+1} \dots u_{l-1})_{k-2} =: (p_1)_{k-2}$ and $M_{y_{l+1} y_\beta} = (u_{l+1} y_{l+2} \dots u_{\beta-1})_{k-2} =: (q_1)_{k-2}$ are determined, as well. With the help of the words p_1 and q_1 the words p_0, q_0 can be expressed in the following way: $p_0 = p_1 y_l$ and $q_0 = y_{l+1} q_1$. Note that if y_{l+1} is a final (or unique) appearance, then q_0 is the empty word. Let us assume that y_{l+1} is a first appearance. If y_{l+1} is not determined uniquely, then the size of its R-block is at least 2, and either y_{l+1} is not the first element of this block or y_{l+1} is the first element of this block, but for this block any permutation π is appropriate. If y_{l+1} is not the first element of this block, then y_l is in the same R-block, therefore $u_l = \emptyset$, so $|u_l| \leq |u_l^*|$. Finally, assume that y_{l+1} is the first element from its R-block, but any π is appropriate for this block. We claim that in this case $q_0 = y_{l+1} q_1 \sim_{k-2} q_1$, so $(q_0)_{k-2}$ is determined, as well. Clearly, $(q_1)_{k-2} \subseteq (q_0)_{k-2}$. Assume that $z = z_1 z_2 \in (q_0)_{k-2}$, where z_1 is the first letter of z. If $z_1 \neq y_{l+1}$, then $z \in (q_1)_{k-2}$ trivially holds. Assume that $z_1 = y_{l+1}$. Clearly, $z_2 \in (q_1)_{k-3}$. As y_β is a final (or unique) appearance, the whole R-block of y_{l+1} is contained in the set $y_{l+1}, y_{l+2}, \dots, y_{\beta-1}$. For any y in the R-block of y_{l+1}, we have that $z_2 y_\beta \in R(y)$. If y_κ is the last letter from the R-block of y_{l+1}, then $z_2 y_\beta \in R(y_\kappa)$. Then for any $y \neq y_{l+1}$ in the R-block of y_{l+1} we have $y z_2 y_\beta \in R(y_{l+1})$. But we know that in this block every ordering of the block-elements would be appropriate, so $y_{l+1} z_2 y_\beta \in R(y_{l+1})$ also holds. As y_β is a final appearance, it follows that $z = y_{l+1} z_2 \in (q_1)_{k-2}$. The dual case of p_0 can be done similarly. Hence, we obtained that $|u_l| \leq |u_l^*|$, since u_l is a shortest solution of $p_0 u_l q_0 \sim r_0$ by induction. There are two special cases: $l = i$ and $l = j$. In these two special cases the definition of p_0 and q_0 is slightly different from the previous definition, but $u_{i,1}$ and $u_{j,2}$ are determined (by p, q and r), so this does not make any difference. □

Remark 12. *Note that the normal form $\overline{w}^{p,q,r}$ is in fact short-lex assuming that we always take the lexicographical order in Step 2 when we are looking for a good permutation π within a block (of course, possibly with the exception of the first element of the block – that is determined with the help of Lemma 7). With this choice, the "block-part" of w (that is, y) is short-lex and for the inner words u_l we might apply induction to see that these are also short-lex.*

As a consequence, the length of the $\sim_k$ normal form is also minimal for any k and any word.

Corollary 13. *Let $k \in \mathbb{N}$. For any word r, the length of the $\sim_k$ normal form of r, denoted by $\hat{r}$, has minimal length.*

Now an upper bound will be given for the maximal possible length of the canonical (so shortest) solution of the equation $pwq \sim_k r$. This will also provide us an upper bound for the index of $\sim_k$.

Proposition 14. *Let $k \in \mathbb{N}$. Let $l_k(n)$ be the maximal possible length of a canonical solution $\overline{w} = \overline{w}^{(p,q,r)}$ if $|c(r)| \leq n$. Then $l_k(n) = \Theta_k(n^{\lceil k/2 \rceil})$. Moreover,* $\max\{|\hat{r}| : |c(r)| \leq n\} = \Theta_k(n^{\lceil k/2 \rceil})$.

Proof. The statement is going to be proved by induction on k. Clearly, $l_1(n) = n$ and $l_2(n) = 2n$. Let $k \geq 3$ and assume that the statement is proved up to $k-1$. According to the definition of $\overline{w}$, we have that $|\overline{w}| \leq 2n + (2n-1)l_{k-2}$, since $|y| = t \leq 2n$, and $u_1, u_2, \ldots, u_{t-1}$ have length at most $l_{k-2}(n)$. By the induction hypothesis $l_k = O_k(n^{\lceil k/2 \rceil})$.

Now, a construction is going to be presented to show that the length of $\hat{r}$ can be $\Omega_k(n^{\lceil k/2 \rceil})$. This construction completes the proof. Let $n_0 = [n/2]$. Let u be a word such that $c(u) \subseteq \{x_1, x_2, \ldots, x_{n_0}\}$ and $|\hat{u}| = \Omega_{k-2}(n_0^{\lceil (k-2)/2 \rceil})$. (According to the induction such u exists.) Let

$$r = x_{n_0+1} u x_{n_0+1} x_{n_0+2} u x_{n_0+2} \ldots x_n u x_n.$$

Clearly, $|c(r)| \leq n$ and $x_{n_0+1} x_{n_0+1} x_{n_0+2} x_{n_0+2} \ldots x_n x_n \leq (\hat{r})'$. For $n_0 + 1 \leq i \leq n$ let w_i be the word formed by the letters between the first and last appearance of x_i in $\hat{r}$. As $w_i \sim_{k-2} u$, the length of the word w_i is at least $|u|$. Therefore, $|\hat{r}| \geq n_0 |u|$, so $|\hat{r}| = \Omega_k(n^{\lceil k/2 \rceil})$.

□

As a corollary, we also get an upper bound for the number of $\sim_k$ equivalence classes, which is denoted by $f_k(n)$ (where n is the size of the alphabet). If each equivalence class contains a word of length at most $O_k(n^{\lceil k/2 \rceil})$, then the number of equivalence classes is at most $n^{O_k(n^{\lceil k/2 \rceil})}$. Hence, $\log f_k(n) = O_k((n^{\lceil k/2 \rceil}) \log n)$. 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. So if k is even the obtained upper bound for $\log f_k(n)$ is tight up to a constant factor. If k is odd, then we have an additional $\log n$ factor, meaning that the shortest representation of some words are $\log n$ times bigger than expected to be based on only the number of different equivalence classes.

4. Algorithm

In this section an algorithm is presented for finding the canonical solution $\overline{w}$ of the equation $pwq \sim_k r$. Let us introduce the notion L for the length of pqr, that is,

$L = |pqr|$, and let n be the size of the alphabet. The input consists of p, q, r and k, while the output is the normal form $\overline{w}^{p,q,r}$. For simplicity we assume that it is possible to read and compare two letters in $O(1)$ steps. The running time of the algorithm will be $O_k((L+n)n^k)$. For each step of the method described in the previous section we are going to present an algorithm with which that step can be done.

If $k = 1$, then $\overline{w}$ is the word containing the letters from the set $C = c(r) \setminus (c(p) \cup c(q))$ exactly once in alpabetical order. The set C can be obtained in $O(L)$ time, and with bin sort the alphabetical order can be determined in $O(L+n)$ time.

Let $k \geq 2$ and assume that the algorithm with the desired running time has been found up to $k-1$.

We start with Step 1. (See Section 3.) The content $c(r)$ can be determined in $O(|r|)$ time. Moreover, it can be also checked which letters appear at least twice, so the multiset $\{y_1, y_2, \ldots, y_t\}$ of the letters of y is determined, and also the U-blocks are found. For each letter a which appears at least twice in r, let r_a be the word obtained from r by deleting all the letters preceeding the first appearance of a in r and the first a. Clearly, a_1 and a_2 are in the same R-block if and only if $r_{a_1} \sim_{k-1} r_{a_2}$, which can be checked with the help of the $\sim_{k-1}$ normal form of the words r_{a_1} and r_{a_2}:

$$a_1 \text{ and } a_2 \text{ are in the same } R\text{-block} \iff \hat{r}_{a_1}^{(k-1)} = \hat{r}_{a_2}^{(k-1)}$$

This way the R-blocks are determined, for instance the R-block of a contains the letters b for which $\hat{r}_a = \hat{r}_b$. The L-blocks can be determined dually. By induction, for a letter a the word $\hat{r}_a$ can be determined in $O((L+n)n^{k-1})$ time, therefore the running time of determining all the blocks is $O((L+n)n^k)$.

By Proposition 2, the elements of each block are consecutive letters in y, and the order of the blocks is determined by r. The order can be obtained by marking the first and last appearances of the letters in r, and taking the corresponding order in y. The running time of this is $O(|r|)$. Now, the blocks and the order of them is determined. We continue with finding C, the content of $\overline{w}$.

The set C was defined as

$$C = \{a \mid \exists v_1, v_2 : v_1 a v_2 \leq r, |v_1 a v_2| \leq k, v_1 a \not\leq p, a v_2 \not\leq q\}.$$

For each letter $a \in c(r)$ we have to decide whether there exist words v_1 and v_2 having total length at most $k-1$ such that $v_1 a \not\leq p$ and $a v_2 \not\leq q$, but $v_1 a v_2 \leq r$. In order to do this, we check for all words w of length at most k, whether they can be written as $w = v_1 a v_2$ in such a way that the above conditions hold. Similarly as before, let q_a be the word obtained from q by deleting all the letters preceeding the first appearance of a in q including the first a. Moreover, dually, let p^a be the word obtained from p by deleting all the letters after the last appearance of a in p including the last a. Note that q_a and p^a can be determined in $O(|p| + |q|)$ time.

Clearly, $v_1 a \not\leq p$ if and only if $v_1 \not\leq p^a$, furthermore, $av_2 \not\leq q$ if and only if $v_2 \not\leq q_a$. Now, our goal is to decide for all words v of length at most $k-1$ whether they can be split into two parts: $v = v_1 v_2$ in such a way that $v_1 a v_2 \leq r$, but $v_1 a \not\leq p$ and $av_2 \not\leq q$. Let $v = w_1 w_2 \dots w_r$, where $|w| = r \leq k-1$. In $O(|p| + |q|)$ time we can determine the smallest index α for which $w_1 w_2 \dots w_\alpha \not\leq p^a$ and the largest index β for which $w_\beta \dots w_r \not\leq q_a$.

Let $r = r_1 r_2 \dots r_{|r|}$. Let us mark the first appearance of w_1 in r (from the left), then after this letter the first appearance of w_2 and so on. For every $1 \leq i \leq |r|$ let $\kappa(i)$ denote the number of marked letters in $r_1 r_2 \dots r_{i-1}$. With other words, $\kappa(i)$ is the unique index for which $w_1 w_2 \dots w_{\kappa(i)} \leq r_1 r_2 \dots r_{i-1}$, but $w_1 w_2 \dots w_{\kappa(i)+1} \not\leq r_1 r_2 \dots r_{i-1}$. Dually, let us mark the last appearance of w_r in r, then preceeding this letter the last appearance of w_{r-1}, and so on. Let $\lambda(i)$ denote the number of marked letters after in $r_{i+1} \dots r_{|r|}$. There exist words v_1 and v_2 such that

$$v = v_1 v_2,\ v_1 a \not\leq p,\ av_2 \not\leq q,\ v_1 a v_2 \leq r$$

if and only if there exists an index $1 \leq i \leq |r|$ such that

$$r_i = a,\ \kappa(i) \geq \alpha,\ \lambda(i) \geq r - (\beta - 1),\ \kappa(i) + \lambda(i) \geq r.$$

For a fixed letter a and a word v of length at most $k-1$ this condition can be checked in $O(L+n)$ time. Hence, the set C can be determined in $O((L+n)n^k)$ time, since there are n choices for a and $O(n^{k-1})$ choices for v.

After finding C, deciding which part of y belongs to p, w and r is straightforward.

At Step 2 and 3 two cases are distinguished: $k = 2$ and $k \geq 3$. We start with the case $k = 2$. In Step 2 an appropriate ordering was defined for each block B. If B is a U-block, then it has only one element, and we are done. If between two first appearances there is no final (or unique) appearance, then they are in the same R-block. So the consecutive first appearances form R-blocks and clearly the lexicographical order is appropriate. The L-blocks can be found similarly, and the lexicographical order is also appropriate. Moreover, Step 3 is not needed when $k = 2$, since $(pwq)' \sim_2 pwq$. Hence, in the case $k = 2$ the overall running time is $O((L+n)n)$.

Now, let $k \geq 3$. We start with Step 2. In this step an appropriate ordering was found for the elements of block B. If B is a U-block, then it has only one element, and we are done. Let us assume that B is an R-block. (The dual case, when B is an L-block can be handled similarly.) The order for B was determined with the help of $(z_v)_{k-1}$, where $(z_v)_{k-1} = (r_b)_{k-1}$ (here $b \in B$ is arbitrary). Therefore, by taking the $\sim_{k-1}$ normal form of at most $2n$ words determined by r, the order for all of the blocks is determined. The running time of this is $O((L+n)n^k)$.

Finally, in Step 3 we have to solve at most $2n-1$ equations over $\sim_{k-2}$. By induction the running time of this is at most $2n-1$ times $O((L+n)n^{k-2})$.

Therefore, the overall running time is at most $O_k((L+n)n^k)$.

Most probably this algorithm can be fastened. However, the (worst-case) running time of such an algorithm is $\Omega(L + n^{\lceil k/2 \rceil})$, since we have to read the words p, q, r and the length of the solution is possibly $\Omega(n^{\lceil k/2 \rceil})$.

5. Acknowledgements

We would like to thank the anonymous referee for the careful reading of the manuscript and useful comments.

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] O. Klíma "Piecewise testable languages via combinatorics on words" *Discrete Mathematics* **311(20)** (2011) 2124–2127.

[6] P. P. Pach "Solving equations under Simons congruence" *Proceedings of the 9th Hungarian-Japanese Symposium on Discrete Mathematics and Its Applications* (2015) 201–206.

[7] J. E. Pin "Varieties of Formal Languages" *North Oxford Academic, Plenum* (1986)

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