On the expressibility of languages by word equations with a bounded number of variables

Juhani Karhumäki [*] Filippo Mignosi [†]
Wojciech Plandowski [‡]

Abstract

A language (resp. a relation) is *expressible* by a word equation e if it is defined as the set of all values of an unknown (resp. a set of unknowns) over all solutions of the equation e. We first present some tools for proving that languages or relations are not expressibles unless a certain number of auxiliary variables enter into the equation. As a consequence an infinite hierarchy - based on the number of auxiliary unknowns - of expressible language is established. Also the necessary number of auxiliary unknowns to encode a boolean combination of word equations into a single equation or inequality is considered. Finally, we present two new tools for establishing the nonexpressibility in general, and, as a consequence, we obtain a gap theorem for expressible languages.

Résumé

Un langage (resp. une relation) est *définissable* par une équation sur les mots e s'il est l'ensemble de toutes les valeurs prise par une variable (resp. un n-uplet de variables) quand on parcourt les solutions de e. On donne des méthodes pour prouver la non-définissabilité de langages ou relations par des equations qui utilisent un nombre fixé de variables auxiliaires. On obtient, comme conséquence, une hierarchie infinie de languages définissable, hierarchie basée sur le nombre de variables auxliaires. On étudie le nombre de variables auxiliaires qui sont necessaires pour coder une combinaison booléenne d'équations sur les mots en une unique equation. On introduit deux nouveaux outils pour prouver la non-définissabilité en géneral. Ces derniers permettent d'établir un "gap theorem" pour les languages définissables.

[*]Supported by Academy of Finland under grant 44087.
[†]Supported by Academy of Finland under grant 44087.
[‡]Supported in part by the grant KBN 8T11C03915 and in part by Academy of Finland under grant 44087.

1 Introduction

Properties of word equations have been studied extensively during the past ten years or so. The seminal Makanin's result, cf. [20], has inspired a lot of research on word equations, see [6] for a survey. Very recently the known complexity of the satisfiability problem was drastically lowered, by showing that the problem is in PSPACE, see [22, 23]. Research on different aspects of word equations was proposed in [13], where the question, originally proposed in [5], was raised about what kinds of languages or relations can be expressed as values of unknowns in solutions of a word equation. Shortly, we would like to know what kinds of properties are expressible by word equations.

It is well known that many simple properties like being a power of a same word, the conjugacy of two words or imprimitiveness of a word are expressible by word equations. The Makanin's result shows that all expressible languages are recursive. Besides that the first result showing the nonexpressibility seems to be that of Büchi, see [3], where he shows that the language $\{a^n b^n : n \geq 1\}$ is not expressible. In [13] a systematic study of expressible and nonexpressible languages was initiated, among other things, several tools to show nonexpressibility was established. However, these tools did not allow to prove that a language cannot be expressed by a certain number of unknowns. A major goal of this note is to fill this gap.

We consider in this paper three problem areas. First in Section 3 we introduce new methods to show that some languages are not expressible. These results are based on subword complexity. Combining this with a method of compressing solutions of word equations, cf. [24], we can prove a gap theorem for expressible languages: the classical subword complexity of an expressible language is either bounded by $\alpha 2^{\log^2 n}$ for some real constant α, or the language must contain a pattern language, cf. [1, 12], and hence the limsup of the complexity is exponential. As a consequence, we can conclude that many sets of power-free words are not expressible.

In Section 4 we consider properties expressible by boolean formulae of equations. As pointed out in [13] each such formula can be transformad into a single word equation expressing the same language or relation. We show here that in this transformation not more than two more unknowns are necessary for any positive formula, ie. for the formula not containing inequality. Further we show that each property expressible by inequality $u \neq v$ using auxiliary unknowns can be expressed by inequality without auxiliary uknowns.

Finally, in Section 5 we prove our main result. We develop a method to show that some expressible languages are no expressible by using only a certain number of auxiliary unknowns. As a conclusion we obtain an infinite proper hierachy of expressible languages based on the number of auxiliary unknowns. Or even more strongly, for each $k \geq 1$, there exists a language expressible by a word equation with $k+1$ unknowns, but not with k unknowns.

2 Preliminaries

We assume that the reader is familiar with the basics of combinatorics of words, cf. [5] Let Σ be an alphabet of constants and Θ be an alphabet of variables. We assume

that these alphabets are disjoint. We use the convention that lower case letters represent constants and capital letters represent variables. We assume an unusual, but for our purposes usefull, convention that positions in a word are between its consecutive letters. Then each word w has $|w|+1$ positions 0,1,...,$|w|$. We say that an occurrence of a subword v of w *overlaps* position i if $w = a_1 \dots a_s v a_k \dots a_{|w|}$ with $a_j \in \Sigma$, and $s < i < k$.

A word equation is a pair of words $(u, v) \in (\Sigma \cup \Theta)^* \times (\Sigma \cup \Theta)^*$ usually denoted by $u = v$. The *size* of an equation e, written as $|e|$, is the sum $|u| + |v|$. A *solution* of a word equation $u = v$ is a morphism $h : (\Sigma \cup \Theta)^* \to \Sigma^*$ such that $h(a) = a$, for $a \in \Sigma$, and $h(u) = h(v)$.

Definition 1. *We say that a language L is* expressible*, if there is an equation e and a variable X such that*

$$L = \{h(X) : h \text{ is a solution of } e\}.$$

Similarly, we say that a relation $\mathcal{R} \in (\Sigma^*)^k$ is expressible *by an equation e if there are variables $X_1, \dots, X_k$ such that*

$$\mathcal{R} = \{(h(X_1), \dots, h(X_k)) : h \text{ is a solution of } e\}.$$

Here we call variables $X_1, \dots, X_k$ the expressing variables *and variables of e from $\Theta - \{X_1, \dots, X_k\}$* auxiliary *varables.*

Hence, in case of expressing a language all variables of the equation except the expressing one are auxiliary.

A *pattern* is a word over alphabets of constants and variables, ie. an element in $(\Sigma \cup \Theta)^+$. A *pattern language* L defined by a pattern p is the set of all morphic images of p under morphisms $h : (\Sigma \cup \Theta)^* \to \Sigma^*$ such that, $h(a) = a$, for each constant $a \in \Sigma$. Assume the pattern p contains t different variables. Then a pattern language defined by p is denoted by $p((\Sigma^*)^t)$. We say that a language L is *pattern-free* if there is no pattern p with $s \geq 1$ variables such that $p((\Sigma^*)^s) \subseteq L$.

A *D0L system* is a triple (Σ, h, w) where Σ is an alphabet, h a morphism $h : \Sigma^* \to \Sigma^*$ and w a word over Σ. The D0L system defines a *D0L language* $\{h^i(w) : i \geq 0\}$ where h^i is an i-folded composition of h, see [26] for more of D0L systems.

3 Tools to prove the nonexpressibility

Let h be a solution of a word equation $e : u = v$ and let $u = u_1 u_2$. A position in $h(u)$ between $h(u_1)$ and $h(u_2)$ is called a *left cut*. Similarly, if $v = v_1 v_2$ a position in $h(u) = h(v)$ between $h(v_1)$ and $h(v_2)$ is called a *right cut*. A *cut* in $h(u)$ is a left or right cut. The proof of the following lemma is anologuous to that of Lemma 6 in [24].

Lemma 2. *Let L be a pattern-free language expressible by a variable X in an equation $e : u = v$. Then for every solution h of e, each subword t of $h(X)$ of length at least two has an occurrence in $h(u)$ overlapping a cut.*

For a word w denote by $p_w(n)$, for $1 \leq n \leq |w|$, the number of different subwords, or factors or contiguous subwords, of w of length n. The function $p_w : N_{|w|} \to N$, where $N_{|w|}$ denotes the set $\{n \in N : n \leq |w|\}$, is called the *subword complexity* of w.

The subword complexity of a language have been determined in a number of cases, see [25] and references therein.

Theorem 3. *Let L be an expressible pattern-free language. Then there is a constant k such that, for each word $w \in L$ and for $1 \leq n \leq |w|$*

$$p_w(n) \leq kn.$$

Proof. Let $w \in L$ and let L be expressible via a variable X in an equation e. Then there is a solution h of e such that $h(X) = w$. By Lemma 2, each subword of w has an occurrence over a cut. The number of cuts is at most $|e|$ so the number of different words of length n which has an occurrence over a cut is at most $|e|n$. Hence, $p_w(n) \leq |e|n$ and it is enough to take $k = |e|$. This completes the proof. ∎

For a language L, denote by $c_L(n)$ the number of words in L of length n. Clearly, for languages which are not pattern-free, ie. contain a pattern language, we have $c_L(an + b) > 2^{\alpha n}$ for some $\alpha > 0$ and integers a, b, i.e. the subsequence $c_L(an + b)$ of $c(n)$ has exponential growth and, consequently, the limsup of the complexity is exponential.

Example 4. Denote by d_n a word that is a fixed concatenation of all words over $\{a, b\}$ of length n containing exactly two occurrences of the letter a and set $L = \{d_n : n \geq 2\}$. Then L is not expressible by word equations. Indeed, since

$$p_{d_n}(n) \geq \begin{pmatrix} n \\ 2 \end{pmatrix}$$

there is no constant k such that for all words w in L, $p_w(n) \leq kn$. Hence, by Theorem 3 L is not expressible by a word equation.

Concerning the subword complexity of a D0L language, the reader can see [5] and the Pansiot's paper mentioned therein.

Example 5. A D0L-language $(\{0, 1, 2\}, h, 0)$ where $h(0) = 012$, $h(1) = 02$, $h(2) = 1$ is not expressible by word equations since its subword complexity is $\Theta(n \log n)$, see [8].

Example 6. Consider the languege

$$L = \{abcbc^2bc^3 \dots bc^n : n \geq 1\}.$$

The language L is generated by D0L-system $(\{a, b, c\}, h, a)$ where $h(a) = abc$, $h(b) = bc$, $h(c) = c$. The subword complexity of the language L is $\Theta(n^2)$, see Example 9.8 in [5]. Hence, L is not expressible by word equations.

In our second tool for proving nonexpressibility we use the fact that words in expressible pattern-free languages are well compressible in terms of special context-free grammars, cf. [24]. Consider a context-free grammar in which each nonterminal occurs on the left hand side of the productions only once which means that for each nonterminal there is at most one derivation which produces a terminal word. The productions in the grammar are of the form $A \to a$, for a terminal symbol a or $A \to B_1[b_1, e_1]B_2[b_2, e_2] \dots B_k[b_k, e_k]$, for $k \leq 3$ where b_i, e_i are positive integers and A, B_i are nonterminals. In the above $B_i[b_i, e_i]$ represents a subword of the terminal word generated by B_i which starts at position b_i and ends at position e_i. The grammar *represents* a terminal word which is derivable from the start nonterminal. The *size* of the grammar is the number of the productions it contains. Let us call grammars of that form *normal*.

Lemma 7. *Let L be an expressible pattern-free language. Let L be a pattern-free language expressible by a word equation of size k. Then for each word $w \in L$ there is a normal grammar of size $k\lceil \log |w| \rceil + 1$ representing w.*

Proof. Let L be expressible via a variable X in an equation $e : u = v$ and let $w \in L$. Then, by Lemma 2, there is a solution h of e such that $h(X) = w$ and each subword of w has an occurrence over a cut in $h(u)$. Let $Cuts$ be the set of all cuts in $h(u)$ and denote $x = h(u) = h(v)$. Consider the words

$$x_{\gamma,i} = x[\max\{\gamma - 2^i, 1\}, \min\{\gamma + 2^i, |x|]$$

for any cut γ and an integer $0 \leq i \leq \lceil \log |w| \rceil$ where the logarithm is at base 2. Clearly, $h(X) = w$ is a subword of one of the words $x_{\alpha,\lceil \log |w| \rceil}$. We have $x_{\gamma,i+1} = sx_{\gamma,i}t$ where $|s|$, $|t| \leq 2^i$ By the definition of h, the word s has an occurrence over a cut, say β. Since $|s| \leq 2^i$ this occurrence is completely contained in, ie. is a subword of, the word $x_{\beta,i}$. Similarly, t is a subword of $x_{\sigma,i}$ for some $\sigma \in Cuts$. This gives the definition of words $x_{\gamma,i+1}$ where γ ranges over $Cuts$ in terms of subwords of the words $x_{\gamma,i}$. These dependences allow to build a normal grammar representing w and containing at most $k\lceil \log |w| \rceil + 1$ productions where k is the size of $Cuts$, in particular $k \leq |e|$. ∎

Theorem 8. *Each expressible pattern-free language L satisfies*

$$c_L(n) < 2^{\alpha 2^{\log^2 n}} \text{ for some real constant } \alpha.$$

Proof. Let L be expressible by a word equation of length k. A grammar of size N represents a word of length at most 3^N. Hence, the indices which occur in the grammar can be represented in $2N$ bits. By Lemma 7, each word in L of length n is represented by a normal grammar of size $k\lceil \log n \rceil + 1$ and therefore can be described by $c \log^2 n$ bits for a suitable constant c. The number of different bit sequences of length at most $c \log^2 n$ is $2^{c \log^2 n + 1} - 2$. This completes the proof. ∎

As a straightforward consequence of Theorem 8 we obtain a gap theorem for languages expressible by word equations.

Corollary 9. *Let L be a language expressible by a word equation. Then either*

$$c_L(n) < 2^{\alpha 2^{\log^2 n}} \text{ for some real constant } \alpha > 0$$

$$\text{or } c_L(an+b) > 2^{\alpha n} \text{ for some } \alpha \text{ and integers } a \text{ and } b,$$

Using above we can re-prove and sharpen some earlier results, cf. [13] and [16].

Example 10. The language $L = (a \cup b)^*$ is not expressible by equations over the alphabet $\{a, b, c\}$ because the language L is pattern-free and $c_L(n) = 2^n$.

Example 11. For $k > 2$ the language of k-power free words is not expressible by word equations since they are pattern-free and the complexity function c_L for them is $2^{\Omega(n)}$, see [5].

We note that it is possible to derive directly from Theorem 3, by using the 1978 version of the Lempel and Ziv compression algorithm (LZ'78), a weaker version of Theorem 8. Indeed next proposition shows a general combinatorial result that links the local complexity of every word of a language to the complexity of the whole language.

Proposition 12. *Let L be a language and let Q be a fixed polynomial such that, for each word $w \in L$ and for $1 \leq n \leq |w|$*

$$p_w(n) \leq Q(n).$$

Then

$$c_L(n) = 2^{o(n)}.$$

Proof. We refer to [15] for notations and definitions not explicitely defined in this proof.

We suppose that the language L is closed by subwords, i.e. it contains as element all subwords (or factors) of words in L. A language closed by subwords is often called in the litterature *factorial*. If L is not closed by subwords, then we can consider its closure by subwords L' (i.e. the language of all subwords of words in L). The hypothesis of the proposition are still verified for L' and the result will hold a fortiori for L.

The classical topological entropy of Languages is defined as $H(L) = \mathit{limsup}\ \frac{1}{n} \log c_L(n)$, where the logarithm is in base 2 and, by convention, $\log 0 = 0$ (cf. [15] and references therein). All we have to prove is that the topological entropy $H(L)$ of L is equal to zero. By Proposition 1 and Proposition 6 of [15] it follows that, in order to prove the proposition, we have to prove that the compression rate $\tau(L) = \mathit{limsup}\ |\gamma(w)|/|w|$ of the LZ'78 compressor γ on L is equal to zero. Therefore it is sufficient to prove that, for $w \in L$, the ratio $|\gamma(w)|/|w|$ tends to zero when $|w|$ tends to infinity. It is known that $|\gamma(w)| \leq m \log m + km$ for some constant k depending on the alphabet, where m is the number of nodes of the Lempel and Ziv trie $T(w)$. The Lempel and Ziv trie $T(w)$ must be a subtrie of the trie T_L representig L, that is the k-ary tree (supposing that $\{1, \cdots k\}$ is the alphabet of L), where the path from the root to each node corresponds to a word in L. Trie T_L is well defined since L is closed by subwords, and hence it is closed by prefixes. The i-th level of trie T_L

contains $p_w(i) \leq Q(i)$ nodes, and therefore the i-th level of the Lempel and Ziv trie $T(w)$ has at most $Q(i)$ nodes.

By very definition of trie $T(w)$, we know that $|w| \geq \Sigma_{i=1}^{m}$ height(v_i), where v's are the nodes of $T(w)$ (cf. [15]). But, since the i-th level of the Lempel and Ziv trie $T(w)$ has at most $Q(i)$ nodes, it follows that

$$\Sigma_{i=1}^{m}\text{height}(v_i) \geq (\Sigma_{i=1}^{J}Q(i)i) + (J+1)r,$$

$0 \leq r \leq Q(J+1)$, where $m = r + (\Sigma_{i=1}^{J}Q(i)$. In other words, in previous inequality we consider a trie with same number of nodes and with "smallest" height, under the constrain imposed by the polynomial Q.

As $|w|$ goes to infinity, the same do m and J. It is known that m is in $\Theta(J^{d(Q)+1})$ and $(\Sigma_{i=1}^{J}Q(i)i)+(J+1)r$ is in $\Theta(J^{d(Q)+2})$, where $d(Q)$ is the degree of the polynomial Q. Therefore m is in $O(|w|^{\frac{d(Q)+1}{d(Q)+2}})$.

By inequality $|\gamma(w)| \leq m \log m + km$, the claim follows. ∎

4 Boolean formulae of equations

In [13] it was shown that all relations or languages expressible by Boolean formulae are actually expressible by a single equation. Here we sharpen this result by paying attention to the number of needed new auxiliary varaibles.

We recall here the notion of expressible language by boolean formulae of word equations. We say that a language L is *expressible by a variable X in a boolean formulae* $\psi(X, X_1, \ldots, X_k)$ on word equations containing variables X, $X_1, \ldots X_k$ if and only if

$$L = \{x : \exists x_1 \ldots \exists x_k \psi(x, x_1, \ldots, x_k)\}.$$

Similarly, we say that a relation $\rho \subseteq (\Sigma^*)^k$ is expressible by variables $Y_1, \ldots, Y_k$ in a boolean formulae $\psi(Y_1, \ldots, Y_k, X_1, \ldots, X_l)$ of word equations containing variables $Y_1, \ldots, Y_k$ and $X_1, \ldots, X_l$ if and only if

$$\rho = \{(y_1, \ldots, y_k) : \exists x_1 \ldots \exists x_l \psi(y_1, \ldots, y_k, x_1, \ldots, x_l)\}.$$

The variables $x_1, \ldots x_l$ are again referred to as auxiliary. We say that a boolean formula is positive if it does not contain *not* operator.

Example 13. The formula $u_1 = v_1$ and $u_2 = v_2$ is equivalent to the equation

$$u_1 a u_2 u_1 b u_2 = v_1 a v_2 v_1 b v_2$$

where a and b are different constants, see [13]. Hence, any relation expressible by a system of equations is expressible by a single equation with the same auxiliary variables.

Lemma 14. *A relation expressible by the formula*

$$\phi : (u_1 = v_1 \text{ or } u_2 = v_2 \text{ or } \ldots u_k = v_k)$$

is expressible by a single equation using only two new auxiliary variables not occurring in ϕ.

Proof. First we prove that we may assume that $u = u_1 = u_2 = \cdots = u_k$. Indeed, the formula ϕ is equivalent to the formula

$$u_1u_2 \dots u_k = v_1u_2 \dots u_k \text{ or } u_1u_2 \dots u_k = u_1v_2u_3 \dots u_k \text{ or } \dots \\ \text{or } u_1u_2 \dots u_k = u_1u_2 \dots u_{k-1}v_k.$$

where left hand sides of all equations in the formula are the same. Define the mapping $<>: (\Sigma \cup \Theta)^+ \to (\Sigma \cup \Theta)^+$

$$< \alpha >= \alpha a \alpha b \text{ with } a, b \in \Sigma, a \neq b.$$

Then, as a consequence of basic results on combinatorics on words, see e.g. [5], for each α, the shortest period of $< \alpha >$ is longer than half of the length of $< \alpha >$, in particular $< \alpha >$ is primitive. Denote $v = v_1v_2 \dots v_k$. Now the result is a consequence of the following equivalence:

$$u = v_1 \text{ or } u = v_2 \text{ or } \dots \text{ or } u = v_k \iff \exists Z, Z' : X = ZYZ', \tag{1}$$

where

$$Y =< v >^2 u < v >^2$$

and

$$X =< v >^2 v_1 < v >^2 v_2 < v >^2 \cdots < v >^2 v_k < v >^2 .$$

The proof of the equivalence (1) follows directly from the fact that the word $< v >^2$ is a prefix and a suffix of Y, and it occurs in X in exactly $k + 1$ places which are indicated in the formula for X. The last fact, in turn, is based on two properties of the word $< v >$. First, since it is primitive it occurs inside the word $< v >^2$ in exactly two places: as a prefix and as a suffix. Second, the word $< v >^2$ cannot occur in $< v > v_i < v >$ since $< v >$ is at least twice as long as v_i, and the shortest period of $< v >$ is longer than the half of its length. ∎

We obtain the following interesting result.

Theorem 15. *Any relation expressible by a positive boolean formula of word equations with k auxiliary variables is expressible by a single word equation with $k + 2$ auxiliary variables.*

Proof. Each positive boolean formula of word equations can be represented as a disjunction of conjunctions. The *and* operator does not require new variables so we first replace all conjunctions by a single word equation. Then, by Lemma 14, the disjunction can be replaced by one equation using two new auxiliary variables. ∎

Example 16. The property "Z is imprimitive" over the alphabet $\{a, b\}$ is expressible by a formula

$$(ZaX = aXZ \text{ or } ZbX = bXZ) \text{ and } (Z = aXaT \text{ or } Z = bXbT).$$

By Theorem 15, this formula can be transformed to a word equation with 5 variables expressing the property. On the other hand, the property is not expressible by a word equation with only one variable. Indeed, suppose it is expressible by a word

equation e with one variable X. Then e is of the form $X \cdots = wX \ldots$. Hence, see [9], all solutions of e are prefixes of the infinite word w^ω. However, there is no infinite word such that all imprimitive words are prefixes of it. The minimal number of needed auxiliary variables is thus something from 1 to 4.

We now turn to consider the expressing power of the inequality.

Theorem 17. *Any relation expressible by a formula* not e *with k auxiliary variables is also expressibile by a formula of the form* not e_1 *without auxiliary variables.*

Proof. Let $\mathcal{R}$ be expressible by variables $X_1, \ldots, X_n$ in an inequality $u_1 \neq u_2$ which contains auxiliary variables $Y_1, Y_2, \ldots Y_k$. We substitute the variables $X_1, \ldots, X_n$ in u_1 and u_2 by any sequence of n words $x_1, \ldots, x_n$ to obtain two words u_1' and u_2' over $Y_1,\ldots Y_k$ and Σ. There are two possibilities. Either the words u_1' and u_2' are identical, and then $(x_1, \ldots, x_n) \notin \mathcal{R}$, or they are not identical and then there is a substitution $Y_1 = y_1$, $Y_2 = y_2 \ldots Y_k = y_k$ such that $u_1' \neq u_2'$ and consequently, $(x_1, \ldots, x_n) \in \mathcal{R}$. If the first case takes place then the variables $Y_1, Y_2, \ldots, Y_k$ occur in u_1 and u_2 in the same order. Moreover, the occurrences of the variables $Y_1,\ldots, Y_k$ in u_1 and u_2 divide these words into two sequences of words over constants and variables $X_1,\ldots, X_n$, and these sequences must be equal under the substitution $X_1 = x_1, \ldots, X_n = x_n$. Conversely, if these conditions are satisfied, then the first case holds true. Hence, the n-tuples which does not belong to $\mathcal{R}$ are precisely those which satisfy a system of equations over variables $X_1,\ldots, X_n$. This system is equivalent to a single equation, say $u = v$, with no new auxiliary variables, see Example 13. Thus, tuples in $\mathcal{R}$ are precisely those which satisfy $u \neq v$. This completes the proof. ∎

We conclude by an interesting consequence.

Corollary 18. *Any relation expressible by a nonequality is a complement of a relation expressible by a word equation without auxiliary variables.*

5 The hierarchy

In this section we show that there exists a proper infinite hierarchy of expressible languages based on the number of auxiliary variables.

Let $\mathcal{S}$ be a set of words of a fixed length. Define a factorization $\mathcal{F}_\mathcal{S}(w)$ of a word w in the following way. If there is no occurrence of a word of $\mathcal{S}$ in w then $\mathcal{F}_\mathcal{S}(w) = w$. Otherwise,

$$\mathcal{F}_\mathcal{S}(w) = w[0, i_1], w[i_1, i_2], \ldots, w[i_k, |w|],$$

where $i_1 < i_2 \cdots < i_k$ are all starting positions of occurrences of words of $\mathcal{S}$ in w.

Theorem 19. *Let L be a pattern-free language which is expressible by word equation e with k variables. Then, for each $i \geq 2$ and for each word $w \in L$ there is a set $\mathcal{S}$ of at most k words of length i such that each factor of $\mathcal{F}_\mathcal{S}(w)$ is shorter than $(|e| + 1)i|e| + i$. Moreover, if $|w| \geq i$, then one of the words in $\mathcal{S}$ is a prefix of w.*

Proof. Take any $w \in L$. Let L be expressible by a variable X in $e : u = v$. By Lemma 2, there is a solution h of e such that $h(X) = w$ and each subword of $h(u) = h(v)$ has an occurrence over a cut. Denote by $\mathcal{S}$ the set of all prefixes of length i of words $h(Y)$ where Y ranges over all variables. It is enough to prove that the factors of $\mathcal{F}_{\mathcal{S}}(h(u))$ are shorter than $(|e|+1)i|e|+i$. Suppose there is a factor f in $\mathcal{F}_{\mathcal{S}}(h(u))$ which is longer than $(|e|+1)i|e|+i$. By the definition of the factorization a prefix of f of length i is in $\mathcal{S}$ and f does not contain an occurrence of a word in $\mathcal{S}$. By Lemma 2, f has an occurrence over a cut. The cut divides the word f into two pieces say f_1 and f_2. Since f does not contain an occurrence of a word in $\mathcal{S}$ we have $|f_2| < i|e|$ (Otherwise it contains a prefix of length i of a word $h(Y)$ for some variable Y). Now, we apply Lemma 2 again but to a factor f_1 being a prefix of f and satisfying $|f_1| > |e|i|e|+i)$. Again f_1 has an occurrence over a cut, and the cut divides the word into two parts say f_1', f_2' such that $|f_2'| < i|e|$. Now $|f_1'| > (|e|-1)i|e|+i$.

We repeat this procedure with f_1' up to the moment when we hit the same cut for the second time. Then we obtain a prefix of f of length at least i which is contained in another prefix of f. This, however, is impossible since f does not contain words in $\mathcal{S}$ as subwords, a contradiction. ∎

As an application of Theorem 19 we consider two concrete examples.

Example 20. Let a, a_1, …, a_k be different letters. The language

$$L = aa_1^*a_2^* \dots a_k^*$$

is expressible by a word equation with $k+1$ variables since it is expressible by the following system of equations:

$$\begin{aligned} a_iX_i &= X_ia_i, \text{ for } 1 \le i \le k \\ Z &= aX_1 \dots X_k \end{aligned}$$

However, L is not expressible by any word equation with k variables. Indeed, suppose it is expressible by a word equation e with k variables. Take $i = 2$, $j = (|e|+2)(2|e|+2)+2$. Then, by Theorem 19, there is a set $\mathcal{S}$ of k words of length 2 such that each factor of $\mathcal{F}_{\mathcal{S}}(aa_1^ja_2^j \dots a_k^j)$ is shorter than $j-2$. This means that there is an occurrence of a word in $\mathcal{S}$ in the subwords a_1^j, a_2^j,…, a_k^j. Hence, all of the words a_1^2, a_2^2,…,a_k^2 are in $\mathcal{S}$. The word aa_1 as a prefix of $aa_1^j \dots a_k^j$ is also in $\mathcal{S}$. Consequently, we have found $k+1$ words of $\mathcal{S}$, a contradiction.

Example 21. The language over two-letter alphabet $\{a, b\}$

$$b(b^ka)^*(b^{k-1}a^2)^* \dots (ba^k)^*$$

is not expressibble by a word equation with k variables. Indeed, assume that it is expressible by an equation e. Take $i = k+1$ and $j = (|e|+2)((k+1)|e|+k+1)+k+1$ and let $\mathcal{S}$ be the set as in Theorem 19. Similarly, as in Example 20 we prove that one of the conjugates of ba^k, one of the conjugates of b^2a^{k-1}, …, and one of the conjugates of b^ka belongs to $\mathcal{S}$, and moreover a prefix b^{k+1} belongs to $\mathcal{S}$. Since

the sets of conjugates are disjoint we again have $k+1$ words in a k element set, a contradiction. On he other hand, the language

$$b(b^k a)^*(b^{k-1}a^2)^* \dots (ba^k)^*$$

clearly is expressible by a word equation with $k+1$ varaibles.

In order to state the main result of this section we denote by $\mathcal{L}_k$ the family of languages which are defined by equations using at most k variables. By Examples 20 and 21, we have:

Theorem 22. *For each $k \geq 1$, $\mathcal{L}_k$ is a proper subset of $\mathcal{L}_{k+1}$. Consequently, there exists an infinite proper hierarchy among expressible languages based on the number of auxiliary variables.*

References

[1] Angluin D., Finding pattern common to a set of strings, *in Proceedings of STOC'79*, 130-141, 1979.

[2] Berstel, J., and Perrin D., *Theory of Codes*, Academic Press, 1985.

[3] Büchi, R. and Senger, S., Coding in the existential theory of concatenation, *Arch. Math. Logik*, **26**, 101-106, 1986/87.

[4] Bulitko, V.K., Equations and inequalities in a free group and a free semigroup, *Tul. Gos. Ped. Inst. Ucen. Zap. Mat. Kafedr. Geometr. i Algebra*, **2**, 242-252, 1970 (in Russian).

[5] Choffrut, C., and Karhumäki, J., Combinatorics of words, *in* G.Rozenberg and A.Salomaa (eds), *Handbook of Formal Languages*, Springer, 1997.

[6] Diekert V., Makanin's algorithm, a Chapter *in* Algebraic aspects of combinatorics on words (Ed.: J. Berstel and D.Perrin), 1999, to appear.

[7] Culik II, K., and Karhumäki, J., Systems of equations and Ehrenfeucht's conjecture, *Discr. Math.*, **43**, 139-153, 1983.

[8] Ehrenfeucht, A. and Rozenberg, G., On the subword complexity of square-free D0L languages, *Theoret. Comput. Sci.* **16**, 25-32.

[9] Eyono Obono, S., Goralcik, P., and Maksimenko, M., Efficient solving of the word equations in one variable, *in Proc. MFCS'94*, LNCS 841, Springer Verlag, 336-341, 1994.

[10] Grigorieff, S., Personal communication.

[11] Harrison, M.A., *Introduction to Formal Language Theory*, Addison-Wesley Publishing Company, 1978.

[12] Jiang T., Salomaa A., Salomaa K., Yu S., Decision problems for patterns, *J. Comput. Syst. Sciences* **50**, 53-63, 1995.

[13] Karhumäki J., Mignosi F., Plandowski W., The expressibility of languages and relations by word equations, J. of the A.C.M., Vol. 47 n.3, 483-505, May 2000.

[14] Khmelevski, Yu. I., Equations in free semigroups, *Trudy Mat. Inst. Steklov*, 107, 1971 (English translation: *Proc. Steklov Inst. of Mathematics 107 (1971)*, American Mathematical Society, 1976.)

[15] Hansel, G., Perrin, D., and Simon, I., Compression and Entropy, *STACS'92*, LNCS n. 577, 515-528, 1992.

[16] Ilie, L., Subwords of power-free words are not expressible by word equations, *Fundamenta Informaticae* **38**, 109-118, 1999.

[17] Koscielski, A., and Pacholski, L., Complexity of Makanin's algorithm, *J. ACM* **43**(4), 670-684, 1996.

[18] Lentin, A., *Equations dans des Monoides Libres*, Gouthiers-Villars, 1972.

[19] Lothaire, M., *Combinatorics on Words*, Addison-Wesley, 1983.

[20] Makanin, G.S., The problem of solvability of equations in a free semigroup, *Mat. Sb.*, Vol. 103,(145), 147-233, 1977. English transl. in *Math. U.S.S.R.* Sb. Vol 32, 1977.

[21] Matijasevich, Y., Enumerable sets are diophantine, *Soviet. Math. Doklady* 11, 354-357, 1970. English transl. in *Dokl. Akad. Nauk SSSR* 191, 279-282, 1971.

[22] Plandowski, W., Satisfiability of word equations with constants is in NEXPTIME, *in:* Proc. STOC'99, 1999.

[23] Plandowski, W., Satisfiability of word equations with constants is in PSPACE, *in:* Proc. FOCS'99, 1999.

[24] Plandowski, W., Rytter W., Application of Lempel-Ziv encodings to the solution of word equations, *in:* Proc. ICALP'98, LNCS 1443, 731-742.

[25] Rozenberg, G., On subwords of formal languages, Lecture Notes in Comp. Science 117 (1981), 328-333.

[26] Rozenberg, G., Salomaa, A., The Mathematical Theory of L systems, Academic Press, 1980.

[27] Seibert, S., Quantifier hierarchies and word relations, Springer LNCS 626, 329-338 (1992).

Juhani Karhumäki
Turku Centre for Computer Science and Department of Mathematics,
Turku University,
20 014, Turku, Finland.
Email:karhumak@cs.utu.fi.

Filippo Mignosi
Dipartimento di Matematica ed Applicazioni,
Universitá di Palermo via Archirafi,
90 123 Palermo, Italy.
Email: mignosi@altair.math.unipa.it.

Wojciech Plandowski Instytut Informatyki, Uniwersytet Warszawski,
Banacha 2, 02–097 Warszawa, Poland
and Turku Centre for Computer Science and Department of Mathematics,
Turku University, 20 014, Turku, Finland.
Email:wojtekpl@mimuw.edu.pl.