UDC 519.76 DOI: 10.18698/1812-3368-2020-3-30-43

ON CERTAIN CLASSES OF IRREGULAR LANGUAGES

A.I. Belousov[1] alexbel@bmstu.ru
R.S. Ismagilov[1] ismagil@bmstu.ru
L.E. Filippova[2] edobrush@gmail.com

[1] **Bauman Moscow State Technical University, Moscow, Russian Federation**
[2] **Higher School of Economics National Research University, Moscow, Russian Federation**

Abstract

Objective of this paper is to prove certain regularity and irregularity conditions in languages determined by a set of integer vectors called distribution vectors of the number of letters in words over a finite alphabet. Each language over the finite alphabet uniquely determines its proprietary set of distribution vectors and vice versa, i.e., each set of vectors is associated with a language having this set of distribution vectors. A single necessary condition for the language regularity was considered associated with the concept of $\mathbb{Z}_+$-plane (sets of points with non-negative integer coordinates lying on a plane in the affine space). The condition is that a set of distribution vectors determined by any regular language could be represented as a finite union of the $\mathbb{Z}_+$-planes. Certain sufficient irregularity conditions associated with the distribution vector properties were proven. Based on this, classes of irregular languages could be identified. These classes are determined by a set of vectors (points) that could not be represented as a finite union of the $\mathbb{Z}_+$-planes; by a set of vectors containing vectors with arbitrarily high values of each coordinate and having certain restrictions on the difference between maximum and minimum values of the coordinates; by a set of vectors called the sparse sets. A method is proposed for building such sets using strictly convex and strictly increasing numerical sequences. These sufficient irregularity conditions are based on the Myhill — Nerode theorem, which is known in the formal languages' theory. Examples of applying the proved theorems to the analysis of languages' regularity/irregularity are presented

Keywords

Formal language, alphabet, regular language, distribution vector, irregularity condition, equivalence relation, equivalence relation index, sparse sets

Received 31.05.2019
Accepted 17.12.2019

Introduction. The problem of analyzing regular languages, including the problem of analyzing regularity/irregularity is very important in the formal languages' theory. This is due to the fact that the regular languages' theory is a linear part of the entire formal languages' theory, generates algorithmic foundation in the syntactic analysis being the basis in the lexical analyzers' development, and also finds other quite non-trivial applications [1]. In addition, publications on purely theoretical terms recently appeared concerning interesting, and partially unexpected properties of regular languages, for example, interconnection between the theory of regular languages and the theory of linear spaces [2]. This work shows mutual reducibility of geometric and linguistic problems.

In general, study of the regular languages' properties remains a topical problem in accordance with several recent publications that reveal new aspects of the long-known concepts [3, 4]. Applications of the regular languages' theory to problems that are not a priori associated with the theory of formal languages are of particular interest [5].

At the same time, questions still remain that were not yet been resolved. These include the study of relationship between certain numerical characteristics of languages and the regularity and irregularity properties.

It should be noted that a direction was being developed for a long time, which could conditionally be called arithmetization of the formal languages' theory, where mutual reducibility of the linguistic problem itself and the problem of the arithmetic functions and relations analysis is taking place (see, for example, [6, 7]).

This article is devoted to the study of certain properties, as well as to establishing certain irregularity sufficient conditions based on the properties of numerical functions and integer vectors. It could be considered as continuation of the work [8], where proof of a single sufficient condition for the languages' irregularity based on the properties of the so-called strongly separable relations on the set of natural numbers was presented. Results obtained provide several new and more efficient tools compared with the known ones in analyzing regularity/irregularity of languages, while giving certain arithmetic and geometric characteristics of regularity and irregularity.

All results concerning irregularity conditions, except for the necessary condition, associated with the concept of $\mathbb{Z}_+$-planes, are based on the Myhill — Nerode theorem known in the theory of formal languages [9, 10]. In accordance with this theorem, criterion for the language regularity is the factor set finiteness determined by a certain equivalence relation associated with language. Voluminous literature is devoted to the Myhill — Nerode theorem (see the literature in Ref. [8]).

Let us mention here only a detailed monograph [11], as well as several works where the theorem generalization for trees is considered [12, 13].

It should be noted that subject of this work is in no way connected with problems related to natural languages, although the authors are aware of several works, where the theory of regular languages is discussed in connection with these problems [14].

Character distribution vectors in language words. Necessary regularity condition. Let us assume that $A = \{a_1, ..., a_n\}$ is the finite alphabet and A^* is the set of all words over this alphabet.

Definition 1. *Let us assume that* $f : A^* \to \mathbb{Z}_+^n$ *is the function associating with each* $u \in A^*$ *word the* $f(u) = (k_1, ..., k_n)$ *vector, where* k_i, $1 \le i \le n$, *is the number of the* a_i *letter occurrences in the* u *word. This vector would be called the character (letter) distribution vector in the a word.*

It should be noted that such function is surjective, i.e., for each α vector the $u \in A^*$ word is defined in such a way that $f(u) = \alpha$.

If L is the language over the A, alphabet, i.e., $L \subseteq A^*$ its image under the function f shall further be denoted by $T(L)$. The $L \mapsto T(L)$ obtained correspondence could be inverted by determining using the T arbitrary set of vectors the $L(T)$, language consisting of all such $u \in A^*$, words, for which $f(u) \in T$.

Let us note that the introduced languages of the $L(T)$ form generate a rather narrow class of languages characterized by the following property: for any $u \in L = L(T)$ word, words obtained from u by any permutation of letters included in it are also contained in the L. Attention to this class is explained by the fact that problem of irregularity in this regard (main problem of this work) is solved simply and naturally. For the arbitrarily given L language, the $L \subseteq L(T(L))$ inclusion is taking place, but not the equality. Obviously, different languages could have the same set of distribution vectors, and the $L(T)$ language is defined as the greatest (under inclusion) language having T, set of distribution vectors.

The described approach leads to the following problems: characterizing the classes of symbol (letter) distribution vectors arising from regular languages and identifying the classes of vectors, for which the corresponding language is irregular.

Definition 2. *Let us denote the set of all sums of the* $f_1 + f_2$ *form, where* $f_1 \in U_1$, $f_2 \in U_2$, *as the sum of the* U_1, $U_2 \subseteq \mathbb{Z}_+^n$ *integer vector sets. This set shell be denoted as* $U_1 \dot{+} U_2$.

Further, having the $T \subseteq \mathbb{Z}_+^n$ set, the $\sum_{i=1}^{N} c_i f_i, f_i \in T,$ set of all finite sums could be considered for the $c_i \in \mathbb{Z}_+$ arbitrary non-negative integer coefficients. Since all coefficients of the linear combination are non-negative integers, this set could be considered as the result of multiple addition of the T set to itself: $T \dot{+} T \dot{+} T \dot{+} \ldots$ Let us denote this set by T^* and call it the T set *iteration*.

Let us recall some definitions from linear algebra.

If S is linear space over the K field, then a plane (affine) in the S is called a set of vectors of the $f_0 + \sum_{i=1}^{N} c_i f_i$ form for the $f_i \in S,\ 0 \le i \le N,$ fixed vectors and the $c_i \in K$ arbitrary elements. Let us provide the following definition.

Definition 3. *Let us call a set of vectors of the* $f_0 + \sum_{i=1}^{N} c_i f_i$ *form for the* $f_0,\ f_i \in \mathbb{Z}_+^n$ *fixed vectors and for the* $c_i \in \mathbb{Z}_+$ *and* $1 \le i \le N$ *arbitrary non-negative integer numbers as the* $\mathbb{Z}_+$*-plane in the* $\mathbb{Z}_+^n$ *set.*

Since vectors involved in the linear combination written above are not required to be linearly independent, different linear combinations could define the same $\mathbb{Z}_+$-plane. Let us call the least of the N numbers for all linear combinations defining the given $\mathbb{Z}_+$-plane as the $\mathbb{Z}_+$*-plane dimension.*

Some comments need to be made regarding the objects entered.

Let us consider the $\mathbb{Z}_+^n$-plane defined by the $f_0,\ f_i, 1 \le i \le N,$ vectors (as explained above), as well as a set in the $\mathbb{R}^n$ (real arithmetic vector space) consisting of all vectors of the $f_0 + \sum_{i=1}^{N} c_i f_i$ form (for the $f_i \in \mathbb{R}^n$ fixed vectors and for the $c_i \in \mathbb{R}$ arbitrary real numbers). This is an affine plane in the $\mathbb{R}^n$ (shift on the f_0 linear span of the system of $f_1, \ldots, f_N$ vectors) containing $\mathbb{Z}_+$-plane and having the same dimension.

Let us consider the $E \subset \mathbb{Z}_+^n$ set to be *finitely generated*, if there exists a finite system of the $f_1, \ldots, f_N,$ vectors, that any $f \in E$ vector could be written as $f = \sum_{i=1}^{N} c_i f_i$ for several $c_i \in \mathbb{Z}_+$ numbers. If T is the finitely generated set of vectors, then the same is the T^* set, and then it is also a plane.

Theorem 1. *If the L language is regular, then the* $T(L)$ *set is the union of finite set of the* $\mathbb{Z}_+$*-planes.*

◀ Theorem proving follows from the lemma below.

Lemma 1. *Let* L_1 *and* L_2 *be the languages over a certain alphabet. Then*

1) $T(L_1 \cup L_2) = T(L_1) \cup T(L_2), T(L_1 L_2) = T(L_1) \dot{+} T(L_2), T(L^*) = (T(L))^*$;

2) let L_1, L_2 *and* L *be the languages over a certain A alphabet, and let each* $T(L_1), T(L_2)$ *and* $T(L)$ *set be the union of a finite set of the* $\mathbb{Z}_+$*-planes. Then, each* $T(L_1) \cup T(L_2), T(L_1) \dot{+} T(L_2), (T(L))^*$ *set is also the union of a finite set of the* $\mathbb{Z}_+$*-planes.*

Proof of the first statement is easily obtained from the fact that when multiplying (concatenating) two words of a language, their distribution vectors are added, and also from taking into consideration definition of language iteration and iteration of a set of vectors.

Second statement is the direct consequence of the first.

Now let us prove the theorem. Any regular language is obtained from the initial elementary regular languages, i.e., from an empty language, the language consisting of a single empty word and from the language that includes one single-character word over the $A = \{a_1, ..., a_n\}$ given alphabet, by using operations of union, concatenation and iteration. Sets of vectors corresponding to these languages are as follows:

1) $T(\varnothing) = \varnothing$;

2) $T(\{\lambda\}) = \overline{0}$ (zero vector in $\mathbb{Z}_+^n$);

3) $T(\{a_i\}) = \{(0, ..., 0, \underset{i}{1}, 0, ..., 0)\}$, where a_i is the character over the $A = \{a_1, ..., a_n\}$, given alphabet.

In 2) and 3) the $\mathbb{Z}_+$-plane of zero dimension is obtained.

It follows from Lemma 1 (second statement) that the sets of distribution vectors corresponding to languages obtained in these operations from the simplest languages are the finite unions of the $\mathbb{Z}_+$-planes. Such is the $T(L)$ set for any regular L language. This proves Theorem 1. ▶

Theorem 1 is true for any regular language, and not just for a language of the $L(T)$ form, but if the T set of vectors is such that it does not satisfy the theorem condition, then any language having this set of distribution vectors would be irregular, including the $L(T)$ greatest (under inclusion) language.

Theorem 1 provides an obvious way to obtain irregular languages: it is sufficient to take the $L(T)$ language, where the T set of vectors could not be represented as a finite union of the $\mathbb{Z}_+$-planes. The simplest example: let the alphabet be $A = \{a_1, a_2\}$, and the T set of vectors would be defined as the set of all vectors of the (m, m^2) form. Then any language having such a set of distribution vectors, including the $L(T)$ language, would be irregular.

The given example could be quite easily analyzed using the pumping (or growth) lemma [15] or the Myhill — Nerode theorem, but a more complicated example gives a language defined by a set of vectors, where the i-th coordinate of each $(k_1, ..., k_n)$ vector is expressed quadratically through the others, namely $k_i = \sum_{j \neq i} c_j k_j^2$ for some $i \in \{1, ..., n\}$ and integer non-negative (at the same time non-zero) c_j numbers (moreover, the set of these numbers depends on the vector).

Irregularity of such language (and together with it of all languages with a given set of distribution vectors) is a direct consequence of Theorem 1. Proving its irregularity using the growth lemma or the Myhill — Nerode theorem requires a rather complicated analysis.

Regarding possibilities of analyzing irregularity using the growth lemma, attention should be paid to [16].

Thus, Theorem 1 turns out to be a very efficient tool in proving irregularity of languages, which, unlike the growth lemma, is not requiring a detailed proper linguistic analysis of any particular language. By virtue of this theorem, it is sufficient to take any set of the T vectors to build an irregular language that could not be represented as the finite union of the $\mathbb{Z}_+$-planes. In particular, such would include the all (infinite) sets of vectors (points) that do not contain any single $\mathbb{Z}_+$-direct (i.e., one-dimensional $\mathbb{Z}_+$-plane). Detailed geometric analysis of such sets is not the subject of this work.

Statement converse to Theorem 1 is not possible. For example, the $\{a^n b^n : n \geq 0\}$ irregular language over the $\{a, b\}$ alphabet defines the $\{(n, n) : n \geq 0\} = \{n(1,1) : n \geq 0\}$ set of vectors generating a single-dimensional plane (right line).

Equivalence of words and equivalence of vectors. Let us consider here only languages of the $L(T)$ form for a certain set of the T vectors.

The Myhill — Nerode theorem [8–10], which is a criterion in the regularity of languages, is based on determining the equivalence relation on a set of words over the A arbitrary finite alphabet that defines the L language over this alphabet.

Definition 4. *Words u and v are considered equivalent in relation to the L language (L equivalent); if for any $x \in A^*$ word, ux and vx words or both belong to the L language, or both do not belong to it, in this case we write $u \equiv_L v$.*

Language regularity criterion mentioned above is as follows: the L language is regular then and only then, if the set of equivalence classes defined by the $\equiv_L$

relation is finite; the number of these classes is called the index of considered equivalence relation.

Next, let us introduce the equivalence relation on the set of $\mathbb{Z}_+^n$ vectors with integer non-negative coordinates. In view of surjectivity of the $f: A^* \to \mathbb{Z}_+^n$ function noted above, each vector of this set could be considered as the distribution vector of the number of characters in a certain word. It is natural to expect that the equivalence relation on a set of vectors is connected to the equivalence relation on the set of words. Such connection is established below.

Definition 5. *Let a set of the $T \subseteq \mathbb{Z}_+^n$ vectors be given. Let us call the $\varphi, \psi \in \mathbb{Z}_+^n$ vectors T-equivalent, if for any $\gamma \in \mathbb{Z}_+^n$ vector, both the $\varphi + \gamma$, $\psi + \gamma$ vectors either belong to the T set or both do not belong to it. Then, let us write: $\varphi \equiv_T \psi$.*

Let us establish connection between the thus determined equivalence of vectors and the equivalence of words defined by a certain language. Let the L language be defined by a set of T vectors, i.e., it consists of all u words over the A alphabet, for which $f(u) \in T$.

Lemma 2. *L-equivalence of the u and v words and is equivalent to the T-equivalence of the $f(u)$ and $f(v)$ vectors.*

◀ Let the u and v words be not L-equivalent: $u \not\equiv_L v$, where the L language is determined by the set of T vectors, i.e., $L = L(T)$. Then, the $x \in A^*$ word could be found, for which the word is $ux \in L$, and the word $vx \notin L$, and consequently $f(ux) = f(u) + f(x) \in T$, $f(vx) = f(v) + f(x) \notin T$. This is followed by $f(u) \not\equiv_T f(v)$. On the opposite, let the $\alpha, \beta \in \mathbb{Z}_+^n$ vectors be non T-equivalent: $\alpha \not\equiv_T \beta$. Then, there appears the γ "separating" vector γ, i.e., $\alpha + \gamma \in T$, $\beta + \gamma \notin T$. Due to surjectivity of the f function there are the $u, v, x \in A^*$, words, for which $f(u) = \alpha$, $f(v) = \beta$, $f(x) = \gamma$, and due to the f function properties we have $f(u) + f(x) = f(ux) \in T$, while $f(v) + f(x) = = f(vx) \notin T$, and, therefore, $ux \in L, vx \notin L$, i.e., the u and v words are not L-equivalent: $u \not\equiv_L v$. ▶

Thus, $L(T)$-equivalence of words is equivalent to the T-equivalence of vectors.

Corollary 1. *The following conditions are equivalent: a) $L(T)$ language is irregular; b) infinite set of pairwise non T-equivalent vectors exist in the $\mathbb{Z}_+^n$ set.*

Note that condition b) is equivalent to the following condition: for any integer r there exists in the $\mathbb{Z}_+^n$ set a subset consisting of r pairwise non T-equivalent vectors.

For further study of equivalence, let us introduce a notation. For any $\varphi \in \mathbb{Z}_+^n$ vector, let us denote by $(T-\varphi)_+$ the set of all $\gamma \in \mathbb{Z}_+^n$ vectors, such as $\gamma+\varphi \in T$. This definition demonstrates that in order to obtain the $(T-\varphi)_+$ set, it is required to take all vectors of the $\alpha-\varphi$, $\alpha \in T$, form and discard vectors (if any) from the resulting set that have at least one negative coordinate.

Lemma 3. *T-equivalence of the* $\varphi, \psi \in \mathbb{Z}_+^n$ *vectors is equivalent to the condition* $(T-\varphi)_+ = (T-\psi)_+$.

◀ Let $\varphi \equiv_T \psi$ and $\alpha \in (T-\varphi)_+$. Then, if $\varphi+\alpha \in T$, and since $\varphi \equiv_T \psi$, then $\psi+\alpha \in T$ and $\alpha \in (T-\psi)_+$, vice versa, if $(T-\varphi)_+ = (T-\psi)_+$, then for any α vector we have $\psi+\alpha \in T \Leftrightarrow \varphi+\alpha \in T$, i.e., $\varphi \equiv_T \psi$. ▶

Thus, the set of vector equivalence classes is identified with the $(T-\varphi)_+$ family of subsets. Hence, if there is an infinite family of (different) $(T-\varphi)_+$ sets, then the $L(T)$ language is irregular.

Irregularity conditions associated with vector oscillation properties. Hereinafter, let us consider certain properties of the set of T vectors, where the $L(T)$ language turns out to be irregular. Let us introduce notation and concepts used below.

For the $\varphi = (\varphi_1, ..., \varphi_n)$, $\psi = (\psi_1, ..., \psi_n)$ vectors, the $\varphi \le \psi$ notation means that $\varphi_i \le \psi_i$ for any $i = 1, ..., n$. Further we assume that $\min\varphi = \min\limits_{1 \le i \le n} \varphi_i$ and $\max\varphi = \max\limits_{1 \le i \le n} \varphi_i$.

Definition 6. *Let us denote the* φ *vector oscillation as the* $\omega(\varphi) = \max\varphi - \min\varphi$ *number. The* $(r, ..., r)$ *vector (with identical components) is denoted by* $\breve{r}$.

Let us also introduce equivalence relation on the $\mathbb{Z}_+$ set of non-negative integers, denoting the *k* and *l* numbers as equivalent, if the $\breve{k}$ and $\breve{l}$ vectors are *T*-equivalent.

Theorem 2. *Let the T set possess the following properties:*

1) for any M number, there is such $\varphi = (\varphi_1, ..., \varphi_n) \in T$, *that* $\varphi_i \ge M$ *for all* $i = 1, ..., n$;

2) there is such $\varepsilon > 0$, *that* $\max\varphi \ge (1+\varepsilon)\min\varphi$ *for any* $\varphi \in T$ *vector.*

Then the $L(T)$ *language is irregular.*

Comment. Property 1) means that the *T* set contains vectors with arbitrarily large values of each coordinate.

Property 2) could be described by the $\omega(\varphi) \ge \varepsilon \min\varphi$ inequality.

Theorem 2 could be proved in the following way: admit the contrary (i.e., the language is regular), apply the required regularity criterion specified in Theorem 1, and arrive at a contradiction (proving that the T set possessing properties (1) and (2) could not be represented as the finite union of planes). However, let us choose here another way based on the Myhill — Nerode theorem.

◀ Assume that the $L(T)$ language is regular. Let us show that in this case the $\mathbb{Z}_+$ set has a pair of equivalent numbers (equivalence relation was introduced before the theorem statement). Indeed, assuming the contrary, we obtain an infinite set of pairwise non T-equivalent vectors of the $\breve{r}$ form that contradicts the $L(T)$ language regularity (by the Myhill — Nerode theorem). Thus, we have a pair of the l,k equivalent numbers. Let us demonstrate that this leads to a contradiction.

It could be assumed that $l>k$. Let us accept that $r=l-k$. Let us take the $\varphi \geq \breve{k}$, vector in the T set, which is possible by virtue of property (1) in the T set. Then, $\varphi-\breve{k} \in (T-\breve{k})_+$. Since the l,k numbers are equivalent, according to Lemma 4, $(T-\breve{k})_+ = (T-\breve{l})_+$, and therefore $\varphi-\breve{k} \in (T-\breve{l})_+$. Consequently, $\varphi-\breve{k}+\breve{l} \in T$, i.e., $\varphi+\breve{r} \in T$. Using a similar argument for the $\varphi+\breve{r}$ vector (instead of the φ vector), the $\varphi+2\breve{r} \in T$ is obtained. Continuing this argument, it could be seen that $\varphi+s\breve{r} \in T$ for any $s \geq 1$. By virtue of property (2), the $\omega(\varphi+sr) \geq \varepsilon \min(\varphi+sr) = \varepsilon sr$ inequity is obtained. Thus, vector oscillation increases unlimitedly (note that ε is positive) in the sequence of $\varphi+s\breve{r} \in T$, $s \geq 1$, vectors. Moreover, each subsequent vector in this sequence is obtained from the previous one by adding a vector with matching coordinates. However, vector oscillation could not alter when adding a vector with matching coordinates. Contradiction. The theorem is proved. ▶

A simple example of a set of the $T \subset \mathbb{Z}_+^n$, $n>1$, vectors satisfying the condition of Theorem 2 is a set consisting of all vectors of the $\varphi = \left(r,(1+p)r,...,(1+(n-1)p)r\right)$ form, where r is the arbitrary non-negative integer; p is the fixed positive integer. It is clear that $\min\varphi = r$, $\omega(\varphi) = (n-1)pr$, $\varepsilon = (n-1)p$. This set defines the $L(T)$ irregular language according to Theorem 2. Obviously, the result would not change under arbitrary permutations of the φ vector coordinates.

In order to derive another sign of irregularity, let us use the following term. Let the infinite X set and the $f: X \to \mathbb{Z}_+^n$ function be provided. Let us call it infinitely great, if $f(x_n) \to \infty$ for any infinite sequence of the $x_n \in X$ pairwise

different elements (this condition is equivale to the following: for any M number, the set of points, where $f(x_n) < M$ is finite).

Theorem 3. *Let the T set of distribution vectors have property (1) from the condition of Theorem 2 and let the $\omega(\varphi)$, $\varphi \in T$, oscillation function be infinitely great. Then, the $L(T)$ language is irregular.*

◀ Let us assume that the $L(T)$ language is regular. Then, as shown in proving Theorem 2, the l, k equivalent numbers are obtained. Repeating the reasoning from the proof of Theorem 2 (this is possible due to property (1)), the $\varphi + s\bar{r} \in T$, $s \geq 1$, infinite sequence of vectors is obtained having the same oscillation (equal to the $\varphi \geq \breve{k}$ vector oscillation). This contradicts the fact that the $\omega(\varphi)$, $\varphi \in T$, function is infinitely great. ▶

A language defined by the $\mathbb{Z}_+^n$ set of vectors in the $(x_m, x_{m+1}, ..., x_{m+n-1})$, $m > 0$, form, where $\{x_m\}_{m>0}$, i.e., Fibonacci sequence, could serve as an example of irregular language satisfying Theorem 3. It should be noted that instead of the Fibonacci sequence any sequence could be taken that is determined by the $x_n = x_{n-1} + x_{n-2}$, $n \geq 2$, recurrence relation under the $x_0 = a_0$, $x_1 = a_1$ arbitrary initial conditions (known as the Lucas sequence). Fibonacci sequence is chosen solely as a specific (and most popular) example.

Let us also note that infinity of the oscillation function does not imply fulfillment of condition 1) of Theorem 2 (existence of arbitrarily large vectors). For example, it is possible to define on the set of two-dimensional vectors a subset, where value of the first component is bounded from above (or even is a constant), and values of the second component are unlimitedly increasing. Besides, if the second component is growing linearly, then the language defined by such set of vectors could quite well turn out to be regular. The simplest example: the $T = \{(2, n) : n \geq 0\}$ set of vectors defines the $L(T) = b^* a b^* a b^* \subset \{a, b\}^*$ regular language.

Sparse sets and irregular languages. Definition 7. *Let us consider the $T \subseteq \mathbb{Z}_+^n$ set to be sparse, if there exists such an infinite sequence of the $\alpha_k \in \mathbb{Z}_+^n$, $k = 1, 2, ...,$ vectors, that the $(T - \alpha_k)_+$ sets are pairwise different.*

Theorem 4. *If $T \subseteq \mathbb{Z}_+^n$ is the sparse set, then the $L(T)$ language is irregular.*

Theorem 4 follows from Lemma 3 and from irregularity criterion that follows from the Myhill — Nerode theorem.

Thus, to obtain irregular languages, it is sufficient to build sparse sets in $\mathbb{Z}_+^n$.

First, let us indicate such sets in $\mathbb{Z}_+$; in this case, we restrict ourselves to simple examples. Let us take the c_k, $k = 1, 2, ...;$ sequence of natural numbers and

denote by S the set of its values (i.e., its range of values as a function of the natural argument).

Theorem 5. *Let the* $c_k, k = 1, 2, \ldots,$ *indicated sequence be strictly increasing and strictly convex (i.e.,* $2c_m < c_{m-1} + c_{m+1}$*). Then the S set is sparse.*

◀ Let us take the c_m number and consider the $(S - c_m)_+ = \{c_k - c_m : k \geq m\}$ set. The smallest positive element of this set is $c_{m+1} - c_m$. Condition of strict convexity results in $2c_m = c_m + c_m < c_{m-1} + c_{m+1}$ followed by $c_m - c_{m-1} < c_{m+1} - c_m$. Thus, the sequence of minima strictly increases. So, these minima are pairwise different; therefore, the $(S - c_m)_+$ sets are pairwise different. ▶

Let us reduce analysis of sparse sets in $\mathbb{Z}_+^n$ to the case of such sets in $\mathbb{Z}_+$. Taking the $E \subset \mathbb{Z}_+^n$ set, let us denote by E_k, $1 \leq k \leq n$, the set compiled of the k-th vector coordinates of the E set. Let us call this set the *k-cut-off* of the E set.

Theorem 6. *If* E_k, $1 \leq k \leq n$, *is a sparse set (for a certain k), then the* $E \subset \mathbb{Z}_+^n$ *set is sparse.*

◀ Let E_k, $1 \leq k \leq n$, be the sparse set. Then, there exists such the countable $c_j, j = 1, 2, \ldots,$ set of numbers that the $(E_k - c_j)_+$ sets are pairwise different. Let us take the $\gamma_j = (0, \ldots, 0, \underbrace{c_j}_{k}, 0, \ldots, 0)$ vectors, where in the j-th vector all coordinates except the i-th are equal to zero, and the k-th coordinate is equal to c_j, and let us consider the $(E - \gamma_j)_+$ sets. Their k-cut-offs are the $(E_k - c_j)_+$ sets that are pairwise different. Therefore, the $(E - \gamma_j)_+$ sets themselves are also pairwise different. Thus, $E \subset \mathbb{Z}_+^n$ is a sparse set. ▶

Indicated above leads to the following simple method for building the sparse sets. Let us select any strictly increasing and strictly convex sequence of numbers. Let us fix the arbitrary $k \in \{1, \ldots, n\}$ and take the E set of vectors, for which the k-cut-off is a set of numbers from the indicated sequence.

A simple example: let us define a sequence of natural numbers, which is a subsequence of Fibonacci numbers starting from 3: 3, 5, 8, 13, 21, 34, ... It is easy to show that it is strictly convex (and, of course, strictly increasing). Then, corresponding sparse set of vectors (of arbitrary dimension) could be determined in such a way that all their coordinates, except for some, could take arbitrary values, and values of several (selected) coordinates generate the above sequence. Language corresponding to this set could easily be determined. In the simplest case of the two-letter alphabet, it is determined by the (m, n_k) set of vectors, where the m number could be anything, and the n_k numbers generate the above subsequence of Fibonacci numbers. It should be noted that in such simple case language irregularity could also be proved using the growth lemma, but the rea-

soning would be much longer and more complicated. Of course, any strictly convex and strictly increasing sequence could be taken, and for each vector component it would be proprietary.

Regarding the examples given in the article, some comments should be made. Due to necessity, these examples are purely illustrative. These are the first "quick" examples that demonstrate possibilities of the proposed theoretical apparatus and are built on the model of well-known similar examples. As a rule, they are connected to the numerical characteristics of words of a language (see examples on growth lemmas in Refs. [15–17], the Myhill — Nerode theorem [8, 11]). Examples in this work are focused primarily on them in order to show higher efficiency of the proposed methods of analyzing regularity/irregularity. In addition, these examples could be of interest in regard to further discussion on connections between the theory of formal languages and other branches of mathematics. Therefore, relations naturally arise with properties of numerical sequences, Fibonacci sequences, in particular, relations between irregular languages and non-linear manifolds in the affine spaces. This is of interest in view of the well-known definition of context-free languages by means of algebraic (non-linear) equation systems in semirings [18]. In the authors' opinion, language irregularity connection with properties of strictly convex sequences, which play an essential role in computational mathematics, appears to be very interesting and rather unexpected.

However, development of real meaningful applications in the approach proposed in this article, for example in theoretical programming, is the subject of a separate publication.

Conclusion. The main results consist in proving some important properties of regular languages in terms of properties of the number of letters distribution vectors in the words of a language. This continues the direction of research presented in Ref. [8], provides some new and more efficient tools in analyzing regularity/irregularity, and also defines certain classes of irregular languages giving certain arithmetic and geometric characteristic of irregularity, which is the main element of this work scientific novelty.

Primarily, Theorem 1 provides necessary regularity condition in terms of the $\mathbb{Z}_+$-planes and determines classes of irregular languages through the manifold properties in affine spaces that could not be represented as the finite union of $\mathbb{Z}_+$-planes. Theorems 2 and 3 connect language irregularity with the set of distribution vectors determining it, where there are vectors with arbitrarily large values of each coordinate and unboundedly growing oscillation. Theorem 4 determines a class of irregular languages in terms of sparse sets and strictly convex numerical sequences that make it possible to build such sets. All these language

analysis tools shed additional light on the possibility of studying the properties of languages based on the Myhill — Nerode theorem. It is equally important that the obtained results provide certain ways in building irregular languages, describe their specific classes in accordance with the distribution vectors properties, and not only help to answer the question of specific languages regularity/irregularity.

In terms of developing the results obtained, it is of interest to generalize such arithmetic and geometric methods of language analysis in relation to their broader classes, and especially important to the context-free languages.

Translated by D.L. Alekhin

REFERENCES

[1] Akhtyorov A.V., Belousov A.I., Voronin A.Yu., et al. Distributed mobile systems for information monitoring. *Information-Measuring and Control Systems*, 2009, no. 6, pp. 27–34 (in Russ.).

[2] Vyalyi M., Tarasov S. Orbits of linear maps and regular languages properties. *J. Appl. Ind. Math.*, 2011, vol. 5, no. 3, art. no. 448.
DOI: https://doi.org/10.1134/S1990478911030173

[3] Denisova D.S. Regular expressions. *Nauchnaya gipoteza*, 2018, no. 5, pp. 37–43 (in Russ.).

[4] Mel'nikov B.F., Vylitok A.A., Mel'nikova E.A. Iterations of languages and finite automata. *International Journal of Open Information Technologies*, 2017, no. 12, pp. 1–17 (in Russ.).

[5] Wang Y., Roohi N., Dullerud G.E., et al. Stability analysis of switched linear systems defined by regular languages. *IEEE Trans. Autom. Control*, 2017, vol. 62, iss. 5, pp. 2568–2575. DOI: https://doi.org/10.1109/TAC.2016.2599930

[6] Ismagilov R.S., Mastikhina A.A., Filippova L.E. On numerical characteristics of formal languages. *Herald of the Bauman Moscow State Technical University, Series Natural Sciences*, 2017, no. 4, pp. 4–15 (in Russ.).
DOI: http://doi.org/10.18698/1812-3368-2017-4-4-15

[7] Allender E., Arvind V., Mahajan M. Arithmetic complexity, Kleene closure, and formal power series. *Theory Comput. Systems*, 2003, vol. 36, no. 4, pp. 303–328.
DOI: https://doi.org/10.1007/s00224-003-1077-7

[8] Belousov A.I., Ismagilov R.S. On one sufficient condition for the irregularity of languages. *Matematika i matematicheskoe modelirovanie* [Mathematics and Mathematical Modeling], 2018, no. 4, pp. 1–11 (in Russ.).
DOI: https://doi.org/10.24108/mathm.0418.0000121

[9] Myhill J. Finite automata and the representation of events. Technical report WADD TR 57-624. Dayton, OH, Wright Patterson Air Force Base, 1957.

[10] Nerode A. Linear automation transformations. *Proc. Amer. Math. Soc.*, 1958, vol. 9, no. 4, pp. 541–544.

[11] Khoussainov B., Nerode A. Automata theory and its applications. In: *Progress in Computer Science and Applied Logic*. Birkhäuser, Boston, Springer, 2001.
DOI: https://doi.org/10.1007/978-1-4612-0171-7

[12] Comon H., Dauchet M., Gilleron R., et al. Tree automata techniques and applications. TATA. 2014. Available at: http://tata.gforge.inria.fr (accessed: 05.04.2019).

[13] Borchardt B. The Myhill — Nerode theorem for recognizable tree series. In: Ésik Z., Fülöp Z., eds. *Developments in Language Theory. DLT 2003. Lecture Notes in Computer Science*, vol. 2710. Berlin, Heidelberg, Springer, 2003.
DOI: https://doi.org/10.1007/3-540-45007-6_11

[14] Grinchenkov D.V., Kushchiy D.N., Spiridonova I.A. [On the implementation of the finite state machine for recognizing formal language descripting partially structured texts]. *Fundamental'nye issledovaniya, metody i algoritmy prikladnoy matematiki v tekhnike, meditsine i ekonomike. Mater. 16-y Mezhdunar. molodezh. nauch.-prakt. konf.* [Fundamental Study, Methods and Algorithms of Applied Mathematics in Technique, Medicine and Economy. Proc. 16th Int. Youth Sci.-Pract. Conf.]. Novocherkassk, Lik Publ., 2017, pp. 49–53 (in Russ.).

[15] Belousov A.I. On presentation technique for certain sections of formal languages theory: pumping lemmas. *Inzhenernyy vestnik* [Engineering Bulletin], 2015, no. 12 (in Russ.). Available at: http://engbul.bmstu.ru/doc/828263.html (accessed: 06.04.2019).

[16] Zhang G.-Q., Canfield E.R. The end of pumping. *Theoretical Comput. Sci.*, 1997, vol. 174, iss. 1-2, pp. 275–279. DOI: https://doi.org/10.1016/S0304-3975(96)00247-2

[17] Aho A.V., Ullman J.D. The theory of parsing, translation, and compiling. Vol. 1. Parsing, Prentice Hall, 1972.

[18] Salomaa A. Formal languages and power series. In: *Handbook of Theoretical Computer Science*. Vol. B. MIT Press, 1990, pp. 103–132.

Belousov A.I. — Cand. Sc. (Phys.-Math.), Assoc. Professor, Department of Mathematical Simulation, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Ismagilov R.S. — Dr. Sc. (Phys.-Math.), Professor, Department of Higher Mathematics, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 1, Moscow, 105005 Russian Federation).

Filippova L.E. — Senior Lecturer, Department of Applied Mathematics, Higher School of Economics National Research University (Myasnitskaya ul. 20, Moscow, 101000 Russian Federation).

Please cite this article as:

Belousov A.I., Ismagilov R.S., Filippova L.E. On certain classes of irregular languages. *Herald of the Bauman Moscow State Technical University, Series Natural Sciences*, 2020, no. 3 (90), pp. 30–43. DOI: https://doi.org/10.18698/1812-3368-2020-3-30-43