ELSEVIER Theoretical Computer Science 290 (2003) 1931–1946

Theoretical Computer Science

www.elsevier.com/locate/tcs

One application of real-valued interpretation of formal power series

An.A. Muchnik*,[1]

Institute of New Technologies, 10 Nizhnyaya Radishewskaya, Moscow 109004, Russia

Received 23 September 2000; received in revised form 16 December 2001; accepted 8 January 2002
Communicated by B. Durand

Abstract

We define two natural properties of context-free grammars. The first property generalizes linearity and the second property strengthens nonlinearity. A language generated by an unambiguous grammar of the first type is called the language with weak linear structure and a language generated by an unambiguous grammar of the second type is called the language with strong nonlinear structure. Our main theorem states that the family of unambiguous grammars generating languages with weak linear structure and the family of unambiguous grammars generating languages with strong nonlinear structure are effectively separable.

Keywords: Context-free grammars; Context-free languages; Unambiguousness; Linearity

1. Introduction

The results given below were announced in [3] and detailed in [4]. We use the method from [6] to define two families of context-free grammars, which are opposite in some sense.

Roughly speaking, a grammar is called weak linear if to each nonterminal we can assign a natural number (called its rank) such that the right-hand side u of each rewriting rule $x \to u$ has no occurrence of a nonterminal with rank greater than rank(x) and has one or no occurrence of a nonterminal with rank equal to rank(x); a grammar is called strong nonlinear if there are arbitrary long derivations and every sufficiently

* Fax: +7-095-915-69-63.

E-mail address: muchnik@lpcs.math.msu.ru (A.A. Muchnik).

[1] The work was partially supported by grant N 01-01-00505 of Russian Foundation for Basic Researches.

PII: S0304-3975(02)00369-9

long derivation includes some nonterminal x such that a word having at least two occurrences of x can be derived from x. Grammars are called "weak linear" and "strong nonlinear" because the first family properly includes the family of all linear grammars and the second family is properly included in the family of all nonlinear grammars.

If we wish to study linearity and nonlinearity not as the properties of grammars but as the properties of languages then it is natural to require the unambiguousness of grammars. For example, the language of all words over the alphabet $\{a, b\}$ can be generated by the linear grammar

$$G = \{x \to xa, x \to xb, x \to a, x \to b\}$$

and by the strong nonlinear grammar $G \cup \{x \to xx\}$. So in the following we shall consider unambiguous grammars only. In some sense unambiguous grammar provides a structure to the generated language (for instance, any language generated by a linear unambiguous grammar has linear structure). Our results are connected with the languages having weak linear or strong nonlinear structure. Our main Theorem 9 states that we can effectively separate the family of unambiguous grammars generating languages with weak linear structure from the family of unambiguous grammars generating languages with strong nonlinear structure. The proof of this theorem uses Semenov's method (see [6]) of substituting reals for terminals in the power series of a grammar. Theorem 6 exhibits some properties of power series of weak linear and strong nonlinear grammars. These properties are used in the main theorem. Finally, we prove that the properties of an unambiguous grammar "to be weak linear" and "to be strong nonlinear" are not invariant. That is, there is an unambiguous grammar that is not weak linear (strong nonlinear) and generates a language with weak linear (strong nonlinear) structure (Theorems 11 and 16).

2. Preliminaries

We begin with some well-known facts from the theory of formal power series. The proofs of them are used in the following.

A context-free grammar is a 4-tuple consisting of:

(1) two disjoint finite alphabets V_N, V_T,

(2) a finite set of rewriting rules,

(3) an initial symbol from V_N.

Symbols from V_N are called *nonterminals*, symbols from V_T are called *terminals*.

Each rewriting rule has the form $x \to u$, where x is a nonterminal and u is a word over $V_N \cup V_T$.

We say that a word w_2 *can be derived in one step from a word* w_1 *by a rewriting rule* $x \to u$ if w_1 and w_2 can be written as follows: $w_1 = v_1 x v_2$, $w_2 = v_1 u v_2$, where v_1, v_2 are words. A sequence $v_1, \dots, v_n$ of words over $V_N \cup V_T$ is called a *derivation of a word* w *from a word* v if $v_1 = v$, $v_n = w$, and for each $i < n$ the word v_{i+1} can be derived in one step from the word v_i by some rewriting rule. Any derivation from the initial nonterminal (one letter word containing the initial nonterminal) is called simply *derivation*.

The grammar *generates a word* w if w can be derived from the initial nonterminal. The set of all words over V_{T} generated by the grammar is called *the language generated by the grammar*.

A grammar is called *reduced* if:

(1) Each nonterminal is a subword of some generated word.

(2) For each nonterminal there is some word over V_{T} derived from it.

(3) The right-hand side of any rewriting rule does not consist of only one nonterminal.

(4) The right-hand side of any rewriting rule is not empty.

The property of a grammar "to be reduced" is obviously decidable. Given a grammar we can construct a reduced grammar that generates the same language except the empty word. (By condition 4 no reduced grammar generates the empty word.)

In the sequel we shall assume that all context-free grammars are reduced (if it is not stated the contrary).

For each grammar let us consider the set of *derivation trees*. A derivation tree is a tree with root; its vertices are marked by symbols from $V_{\mathrm{N}} \cup V_{\mathrm{T}}$. The sons of each vertex are linearly ordered "from left to right". All the descendants of "more left" brother are considered to be more left than the descendants of all "more right" brothers. The notion of a *derivation tree from a nonterminal* is defined by induction.

Base of induction. A tree that contains only the root, which is marked by a nonterminal, is a derivation tree from this nonterminal.

Induction step. If R is a derivation tree from a nonterminal y, l is a leaf marked by a nonterminal x, and the grammar contains a rewriting rule $x \to a_1 \dots a_n$, then the tree obtained from R by adding n edges with begins in l and with ends marked by $a_1, \dots, a_n$ from left to right is a derivation tree from y.

The size of a derivation tree is defined as the number of branching vertices. If a tree consists of the root only then the root is considered as a leaf. We say that R *is a derivation tree* of a word u if u can be obtained by reading the marks on the leaves of R from left to right. Derivation trees from the initial nonterminal are called simply *derivation trees*. Obviously, a word is generated iff it has a derivation tree.

Lemma 1. *For any reduced grammar the size of any derivation tree of a word* w *is not more than* $2\,\mathrm{length}(w) - 1$.

Proof. By induction. If the statement of the lemma holds for all subtrees with roots in the sons of the root of a derivation tree, then it is easily shown that the statement holds for the whole tree.

To formulate the first theorem we need the notion of a formal power series. *A formal power series* (FPS), or simply a power series, of (noncommuting) variables $t_1, \dots, t_m$ is a formal sum $a = \sum_w w(a, w)$, where w ranges over the set of all words over the alphabet $\{t_1, \dots, t_m\}$, and $(a, w) \in \mathbb{R}$. We shall say that (a, w) is the *coefficient of a on w*. We define the sum of power series according to the rule

$$\sum_w w(a,w) + \sum_w w(b,w) = \sum_w w((a,w) + (b,w))$$

and the product of FPS according to the rule

$$\left(\sum_w w(a,w)\right)\left(\sum_w w(b,w)\right)=\sum_w\left(w\sum_{uv=w}(a,u)\cdot(b,v)\right)$$

(the notation $uv=w$ means that the concatenation of the words u and v is equal to w).

We shall consider only series $\sum_w w(a,w)$ with zero constant term, i.e., the series such that $(a,\Lambda)=0$, where Λ stands for the empty word. The sum and the product of such series have zero constant terms too.

Following [6], to each grammar assign FPS $\sum_w w(a,w)$, where variables are terminals of the grammar and (a,w) is equal to the number of different derivation trees of the word w (this number is finite due to Lemma 1). If the grammar does not generate a word w, then $(a,w)=0$. Since the grammar is reduced, its FPS has zero constant term. We shall say that this FPS is the *formal power series of the grammar*.

Semenov suggested [6] (see also [5]) to consider the FPS of a grammar as usual power series in the sense of analysis and to substitute positive real numbers for the terminals. Let us sketch this method. The FPS of a grammar can be obtained from solution of some algebraic system of equations, where unknowns range the set of all FPS. If we substitute positive reals from the domain of FPS' convergence for terminals in FPS and the system of equations, then the sum of obtained numerical series satisfies the obtained numerical system of equations.

The following well-known theorem (see [5]) formalizes the previous statements. Its proof will be used in the sequel.

First let us introduce some notation. Let r be a power series in variables $t_1,\dots,t_m$, $x_1,\dots,x_n$. Let $s_1,\dots,s_n$ be formal power series. Let us substitute $s_1,\dots,s_n$ for $x_1,\dots,x_n$ in r. Since $s_1,\dots,s_n$ have no constant term (recall that we consider FPS without constant term only), the result is a new power series. Denote it by $r[s_1,\dots,s_n]$.

Let $r_1,\dots,r_n$ be power series in variables $t_1,\dots,t_m$, $x_1,\dots,x_n$. Consider the equalities

$$x_i=r_i,\quad i=1,\dots,n \tag{1}$$

as the system of equations with unknowns $x_1,\dots,x_n$. We say that a *solution* of (1) is an n-tuple $\langle s_1,\dots,s_n\rangle$ of FPS in variables $t_1,\dots,t_m$ if $s_i=r_i[s_1,\dots,s_n]$ for all $i=1,\dots,n$. The main example of a system of this kind is the following. Suppose that we have a (reduced) grammar with terminals $t_1,\dots,t_m$ and nonterminals $x_1,\dots,x_n$. Let $x_i\to u_1^i,\dots,$ $x_i\to u_{k_i}^i$ $(i=1,\dots,n)$[2] be all rules with the left-hand side x_i. Evidently, the system

$$x_i=u_1^i+\dots+u_{k_i}^i,\quad i=1,\dots,n \tag{2}$$

has the form (1).

Theorem 2. I. *Let for all i,j no r_i has terms of the form ax_j, $a\neq 0$. Then system* (1) *has a unique solution.*

II. *If $\langle s_1,\dots,s_n\rangle$ is the solution of system* (2) (*unique according to part* I), *then (s_i,w) is equal to the number of derivation trees of w from x_i (for all $w\in\{t_1,\dots,t_m\}^*$ and for all $i=1,\dots,n$).*

[2] The symbol i in u^i is an index, not an exponent.

Proof. Let $i \in \{1,\dots,n\}$ and let $s_1,\dots,s_n$ be power series (with zero constant term). As r_i has no terms of the form ax_j, $a \neq 0$, coefficient of the series $r_i[s_1,\dots,s_n]$ on any word w depends only on coefficients of $s_1,\dots,s_n$ on the words of length less than $|w|$ (by $|w|$ we denote the length of w). Thus for all w we have

$$\begin{aligned}&\text{if } (s_i,v) = (u_i,v) \text{ for all } i \text{ and for all } v \text{ such that } |v| < |w| \\ &\quad\text{then } (r_i[s_1,\dots,s_n],w) = (r_i[u_1,\dots,u_n],w) \text{ for all } i.\end{aligned} \tag{3}$$

Now let us prove the part I of the theorem.

First let us prove the uniqueness of the solution of (1). Assume that power series $s_1,\dots,s_n$; $u_1,\dots,u_n$ of variables $t_1,\dots,t_m$ satisfy (1), that is,

$$s_i = r_i[s_1,\dots,s_n], \quad i = 1,\dots,n \tag{4}$$

and

$$u_i = r_i[u_1,\dots,u_n], \quad i = 1,\dots,n. \tag{5}$$

Let us prove that $\forall i\ s_i = u_i$. Assume the contrary: $\exists i\ s_i \neq u_i$. Denote by w one of the shortest words over $\{t_1,\dots,t_m\}$ such that there is $i \in \{1,\dots,n\}$ such that $(s_i,w) \neq (u_i,w)$. Then for all words v of length less than $|w|$ and for all i we have $(s_i,v) = (u_i,v)$. Implication (3) combined with (4) and (5) yields that $(s_i,w) = (u_i,w)$. Contradiction.

Secondly, let us prove the existence of solution of system (1). We define a sequence of n-tuples $\langle s_1^\alpha,\dots,s_n^\alpha\rangle$, $\alpha \in \mathbb{N}$, where $s_1^\alpha,\dots,s_n^\alpha$ are power series of variables $t_1,\dots,t_m$, by induction on α.

Base of induction. $s_1^0 = \cdots = s_n^0 = 0$

Step of induction. For each i

$$s_i^{\alpha+1} = r_i[s_1^\alpha,\dots,s_n^\alpha].$$

We claim that for each $w \in \{t_1,\dots,t_m\}^*$, each $i \in \{1,\dots,n\}$, and each $\alpha \geqslant |w|$, $\beta \geqslant |w|$ it holds

$$(s_i^\alpha,w) = (s_i^\beta,w).$$

We prove this by induction on $|w|$. If $w = \Lambda$, then $(s_i^\alpha,w) = 0$ for all $\alpha \in \mathbb{N}$, because r_i has no constant term. Assume that $w \in \{t_1,\dots,t_m\}^*$, $|w| \geqslant 1$ and our assertion holds for every v such that $|v| < |w|$. Take arbitrary $\alpha \geqslant |w|$, $\beta \geqslant |w|$. By induction hypothesis $(s_i^{\alpha-1},v) = (s_i^{\beta-1},v)$ for all v such that $|v| < |w|$ and all i. Thus (3) involves that $(s_i^\alpha,w) = (s_i^\beta,w)$ for all i.

Let $\langle s_1,\dots,s_n\rangle$ be the “limit” of the sequence $\langle s_1^\alpha,\dots,s_n^\alpha\rangle$. More precisely, (s_i,w) is the common value of (s_i^α,w), when $\alpha \geqslant |w|$. Let us prove that $\langle s_1,\dots,s_n\rangle$ is a solution of (1), that is, $s_i = r_i[s_1,\dots,s_n]$, $i = 1,\dots,n$. Take arbitrary $w \in \{t_1,\dots,t_m\}^*$ and prove that $(s_i,w) = (r_i[s_1,\dots,s_n],w)$. Let us pick some $\alpha \geqslant |w|$. Then

$$(s_i,w) = (s_i^\alpha,w) = (r_i[s_1^{\alpha-1},\dots,s_n^{\alpha-1}],w).$$

Using (3), we get

$$(r_i[s_1^{\alpha-1},\ldots,s_n^{\alpha-1}],w)=(r_i[s_1,\ldots,s_n],w).$$

This completes the proof of part I.

Let us prove part II. Denote by r_i the power series $u_1^i+\cdots+u_{k_i}^i$. As the grammar is reduced, system (2) satisfies the assumptions of part I.

Define the sequence of n-tuples $\langle p_1^\alpha,\ldots,p_n^\alpha\rangle$, where p_i^α is a power series in variables $t_1,\ldots,t_m$, $x_1,\ldots,x_n$, by the equalities $p_i^0=x_i$, $p_i^{\alpha+1}=r_i[p_1^\alpha,\ldots,p_n^\alpha]$. And define also a corresponding sequence of derivation trees: each tree of level 0 is a single vertex marked by a nonterminal (x_i); each tree of level $(\alpha+1)$ is obtained from a tree of level α by applying some rules to all nonterminals on the leaves. It is easy to prove (by induction) that $(p_i^\alpha,w)=$(number of derivation trees of w from x_i of level α) for any word $w\in(V_\mathrm{N}\cup V_\mathrm{T})^*$, and if a tree of level α derives a word containing nonterminals, then length of w is greater than α. Hence if $w\in\{t_1,\ldots,t_m\}^*$, then level $\alpha=|w|$ contains all derivation trees of w. Comparing the definitions of p_i^α and s_i^α (s_i^α and s_i are defined in the proof of part I), we see that $s_i^\alpha=p_i^\alpha[0,\ldots,0]$. Therefore for all $w\in\{t_1,\ldots,t_m\}^*$ it holds $(s_i^\alpha,w)=(p_i^\alpha,w)$. Let $\alpha=|w|$. Then $(s_i,w)=(s_i^\alpha,w)=(p_i^\alpha,w)=$(number of derivation trees of w from x_i of level α)$=$(number of derivation trees of w from x_i).

Theorem 2 is proved. □

3. Weak linear and strong nonlinear grammars: definition and decidability

Recall that a grammar is called *linear* if the right-hand side of any rewriting rule has not more than one occurrence of a nonterminal. We say that a nonterminal x is *nonlinear* if some word with at least two occurrences of x can be derived from x.

Definition 3. A grammar is called *weak linear* if it does not have nonlinear nonterminals.

We say that a (finite) *quasiderivation* is any finite sequence $x_1,x_2,\ldots,x_n$ of nonterminals such that x_1 is the initial nonterminal and for every $i<n$ there is a rule $x_i\to u$ such that u has an occurrence of x_{i+1}.

Definition 4. A grammar is *strong nonlinear* if (1) there are arbitrary long quasiderivations and (2) every sufficiently long quasiderivation has a nonlinear nonterminal.

An *infinite quasiderivation* is any infinite sequence $x_1,x_2,\ldots,x_i,\ldots$ if all beginnings of it are finite quasiderivations.

Due to Koenig's lemma we get an equivalent definition of strong nonlinearity if we say that there is an infinite quasiderivation and every infinite quasiderivation has a nonlinear nonterminal.

Obviously, every strong nonlinear grammar generates an infinite language.

Theorem 5. *Both properties of a grammar, to be weak linear and to be strong nonlinear, are decidable.*

Proof. To prove that the property "to be a weak linear grammar" is decidable it is sufficient to prove that the property "to be a nonlinear nonterminal" is decidable. To this end for a nonlinear nonterminal x we estimate the minimal size of derivation tree from x having (at least) two leaves marked by x. Obviously, the size is not greater than $N_{\mathrm{s}}^{L_{\mathrm{b}}}$, where N_{s} is the maximum number of sons of a vertex, L_{b} is the maximum length of a branch. The number of sons of each vertex is bounded by the maximal length of the right-hand sides of rewriting rules. Let us fix two different branches π_1 and π_2 of such derivation tree beginning in the root and ending in leaves marked by x. Suppose that the length of some branch π of this derivation tree is greater than the tripled number of nonterminals. Let us divide this branch into three parts: (1) the part of π common with both π_1 and π_2, (2) the part of π common with only one branch π_1 or π_2 and (3) the remaining part of π. There is a part, whose length is greater than the number of nonterminals. There is a nonterminal having two occurrences in this part (say in vertices γ and δ). Let us replace the subtree, whose root is γ, with the subtree, whose root is δ. The new derivation tree again has two leaves marked by x and its root is also marked by x. If γ is lower in the tree than δ, then size of the new tree is less than size of the old one.

Let us prove the decidability of the second property. To this end we shall prove that both items in the definition of strong nonlinearity are decidable.

Evidently, there are arbitrary long quasiderivations iff there is a quasiderivation having two occurrences of the same nonterminal. If such a quasiderivation exists, then there is such a quasiderivation with the length not more than the number of nonterminals plus 1. Therefore, we can decide whether there is such a quasiderivation.

In the same way, we can prove that there are arbitrary long quasiderivations without nonlinear nonterminal iff there is a quasiderivation with length not more than the number of nonterminals plus 1 having two occurrences of the same nonterminal and having no nonlinear nonterminal. Therefore, the second item of the definition of a strong nonlinear grammar is also decidable.

Theorem 5 is proved. □

Obviously, linearity implies weak linearity and weak linearity is inconsistent with strong nonlinearity.

4. Power series of weak linear and strong nonlinear grammars

The following theorem is an important tool for our further studies. In its formulation terminals are considered to range the set of positive real numbers.

Theorem 6. I. *For any weak linear grammar the domain of its FPS convergence is open.*

II. *For any strong nonlinear grammar the domain of its FPS divergence is open.*

Proof. I. Let us be given a weak linear grammar.

We say that a nonterminal y *follows* a nonterminal x if some word that contains y can be derived from x. Two nonterminals are called *equivalent* if they follow each other.

Let us define rank of nonterminals. For each $r \in \mathbb{N}$ define the set A_r of nonterminals by induction: $A_0 = \emptyset$, $A_{r+1} =$ (the set of all nonterminals x such that any nonterminal following x belongs to A_r or is equivalent to x). Evidently, $A_0 \subset A_1 \subset A_2 \subset \cdots$ and for every nonterminal x there is r such that $x \in A_r$. The *rank* of a nonterminal x is the least r such that $x \in A_r$.

The reader can easily verify the following properties of the rank:

(1) If the left-hand side of some rewriting rule contains a nonterminal of rank r, then the right-hand side does not contain nonterminals of rank more than r.

(2) If the left-hand side of some rewriting rule contains a nonterminal x and the right-hand side contains a nonterminal y of the same rank, then x and y are equivalent.

A weak linear grammar has one more property.

(3) If the left-hand side of some rewriting rule is a nonterminal of rank r, then its right-hand side cannot have the form uy_1vy_2w, where y_1 and y_2 are (possibly equal) nonterminals of rank r.

We can define a weak linear grammar as a grammar satisfying the third property. The new definition is equivalent to the definition given in Section 3.

We must prove that the domain of convergence of grammar's power series is open. According to Theorem 2 the power series of the grammar satisfies system (2).

In the case of a linear grammar system (2) consists of equations linearly depending on nonterminals. In the case of a weak linear grammar system (2) can be split into "almost linear" systems. Namely, let us separately consider the system consisting of all equations with nonterminals of rank r in the left-hand side; substitute certain power series for the nonterminals of rank less than r; then we get a linear system.

Let us solve system (2) in the following way. First let us solve the system consisting of all equations with the left-hand side having rank 1. Secondly, let us solve the system consisting of all equations with the left-hand side having rank 2. This system is also linear system (because we know the values of all nonterminals of rank 1), its coefficients include power series found in the previous step for nonterminals of rank 1. And so on.

Let us call a power series *good* if the following three conditions hold (recall that we consider only series with zero constant term and the range of variables is the set of positive reals):

(1) all coefficients of the series are nonnegative;

(2) the domain of convergence of the series can be specified by some system of strict rational inequalities (in other words, inequalities of the form $f(t_1,\ldots,t_m)/g(t_1,\ldots,t_m) > 0$, where f and g are polynomials with integer coefficients); and

(3) within the domain of convergence the sum of the series is a rational function of its variables.

Note that any series consisting of a single variable is good.

Lemma 7. *Let*

$$\left\{x_i = b_i + \sum_{j=1}^{n} a_{ij}x_j\colon\ i = 1,\ldots,n\right\} \tag{6}$$

be a system in unknown FPS $x_1,\ldots,x_n$ and all b_i and a_{ij} be good FPS. Then this system has a unique solution $\langle s_1,\ldots,s_n\rangle$, where all s_i are good FPS.

Proof. From Theorem 2 it follows that system (6) has a unique solution. Moreover the well-known theorem from [5] states that solutions s_i of system (6) can be obtained from its coefficients with a finite number of applying the operations of sum, product and the binary operation $p,q \to p^* \cdot q$. The operation s^* (or iteration) can be applied to FPS with zero constant term only and is defined by the equality $p^* = 1 + p + p^2 + \cdots$. Note that p^* has nonzero constant term but $p^* \cdot q$ has zero constant term (if q has zero constant term).

This theorem is a direct generalization of the well-known theorem of Kleene on regularity of recognizable sets.

We must prove that $s_1,\ldots,s_n$ are good series. Let us prove that if p and q are good series, then $p+q$, $p\cdot q$, $p^*\cdot q$ are again good series. The validity of condition (1) for $p+q$, $p\cdot q$, $p^*\cdot q$ is evident. Let us verify that $p+q$, $p\cdot q$, $p^*\cdot q$ satisfy (2) and (3).

Sum. The domain of convergence of the series $p+q$ is equal to the intersection of the domains of convergence of p and q because due to (1) all terms are nonnegative. The intersection of the domains corresponds to the union of systems defining these domains.

The sum of series $p+q$ is equal to the sum of series p plus the sum of series q, consequently this sum is again a rational function.

Product. If one of series is identically zero, then the validity of (2) and (3) is evident.

In the other case the domain of convergence is again the intersection of the domains of convergence and the sum of series $p\cdot q$ is equal to the product of the sum of series p and the sum of series q.

Iteration. The domain of convergence of series r^* can be defined by the system defining the domain of convergence of r and the additional inequality $\mathrm{sum}(r) < 1$.

The sum of series r^* is equal to $1/(1-\mathrm{sum}(r))$.

Lemma 7 is proved. □

Note that if we substitute reals for variables of FPS, then the result of substituting does not depend on the order of multipliers in terms.

Now recall that we solved system (2) sequentially and considered the subsystems of (2) corresponding to a certain rank of the left-hand side. Let us prove by induction that all these subsystems satisfy the assumptions of Lemma 7. Every subsystem can be reduced to the form (6) by permutating multipliers in terms. And the coefficients b_i, a_{ij} are obtained from the terminals and the solutions of the previous subsystems by additions and multiplications. Note that the solutions of the new system as functions from $\mathbb{R}^m$ into $\mathbb{R}$ are equal to the solutions of the old system.

To complete the proof in the weak linear case it suffices to note that any set defined by a system of strict rational inequalities is open.

II. Suppose we have a strong nonlinear grammar.

Consider system (2) associated with this grammar. Let $\langle s_1,\dots,s_n\rangle$ be the solution of (2). Recall that s_i is FPS of variables $t_1,\dots,t_m$, where $t_1,\dots,t_m$ are the terminals of the grammar. If $c_1,\dots,c_m$ are positive reals, then by $s_i(c_1,\dots,c_m)$ we denote the result of substituting $c_1,\dots,c_m$ for $t_1,\dots,t_m$ in s_i (so $s_i(c_1,\dots,c_m)$ is a numerical series).

Let x_1 be the initial nonterminal and let $(c_1,\dots,c_m)$ belong to the domain of divergence of the series s_1.

The following assertions will be useful.

(1) There exists a nonlinear nonterminal x_i such that the numerical series $s_i(c_1,\dots,c_m)$ diverges.

(2) For every nonlinear nonterminal x_i the domain of divergence of the series s_i is open.

Let us deduce from these assertions that there is a neighborhood of the point $(c_1,\dots,c_m)$ such that the series s_1 diverges in all points of this neighborhood.

Let us fix some nonlinear nonterminal x_i such that s_i diverges in $(c_1,\dots,c_m)$. Because x_i follows x_1, there is $\alpha\in\mathbb{N}$ such that some term of p_1^α has an occurrence of x_i (the power series p_j^β, $\beta\in\mathbb{N}$, $j=1,\dots,n$, were defined in the proof of part II of Theorem 2 by equalities

$$p_j^0=x_j,\quad p_j^{\beta+1}=r_j[p_1^\beta,\dots,p_n^\beta],$$

where r_j stands for the right-hand side of the jth equation in (2)). Using induction, we can easily prove that for all $\beta\in\mathbb{N}$ and for all $j=1,\dots,n$ it holds $s_j=p_j^\beta[s_1,\dots,s_n]$.

Hence $s_1=p_1^\alpha[s_1\dots s_n]$. Condition (2) in the definition of reduced grammar and Theorem 2 (part II) imply that each power series s_j has a nonzero term. Therefore, its sum is positive in every point $(d_1,\dots,d_m)$ such that $d_1>0,\dots,d_m>0$. The series s_i occurs to some nonzero term of $p_1^\alpha[s_1\dots s_n]$, consequently $p_1^\alpha[s_1\dots s_n](=s_1)$ diverges in every point, where s_i diverges.

Thus it remains to prove assertions (1) and (2).

Let us prove that there is a nonlinear nonterminal x_i such that the series $s_i(c_1,\dots,c_m)$ diverges. Due to the strong nonlinearity it is sufficient to construct an infinite quasiderivation such that every nonterminal in it has divergent series in the point $(c_1,\dots,c_m)$. This quasiderivation begins with the initial nonterminal x_1. System (2) yields $s_1=u_1^1[s_1\dots s_n]+\dots+u_{k_1}^1[s_1\dots s_n]$. Thus there is j such that $u_j^1[s_1\dots s_n]$ diverges in $(c_1,\dots,c_m)$. Further, u_j^1 contains a nonterminal, whose series diverges in $(c_1,\dots,c_m)$. Let us apply the same reasoning to this nonterminal (say x_2). Repeating this procedure, we get the desired quasiderivation.

Let us now prove that the domain of divergence of each nonlinear nonterminal is open. Let x_i be a nonlinear nonterminal and let s_i diverges in a point $(c_1,\dots,c_m)$. By definition of the nonlinearity there is α such that the series p_i^α has a term having at least two occurrences of x_i. As mentioned above, $s_i=p_i^\alpha[s_1\dots s_n]$ (for all α). After some permutation of multipliers we get $s_i=s_i^2\cdot p[s_1\dots s_n]+q[s_1\dots s_n]$, where p and q are power series with integer coefficients and p is not identically zero. Let σ be some positive number less than the sum of the series $p[s_1 \dots s_n]$ in the point $c_1, \dots, c_m$ (the sum of the divergent series is considered to be equal to $+\infty$). It is easy to see that there is a neighborhood U of $(c_1 \dots c_m)$ such that in all its points $p[s_1 \dots s_n]$ is greater than $\sigma/2$. As the series s_i diverges in $(c_1, \dots, c_m)$, there is a neighborhood V of $(c_1 \dots c_m)$ such that in all its points the sum of s_i is greater than $4/\sigma$. Let us prove that in all points in $U \cap V$ the series s_i diverges. Let us take some point in $U \cap V$ and denote by s^+ sum of a series s in this point. Then $s_i^+ = (s_i^+)^2 p[s_1 \dots s_n]^+ + q[s_1 \dots s_n]^+$. From $p[s_1 \dots s_m]^+ > \sigma/2$ and $s_i^+ > 4/\sigma$ we can deduce that $s_i^+ > (4/\sigma)^2(\sigma/2) = 8/\sigma$. Therefore, $s_i^+ > (8/\sigma)^2(\sigma/2) = 32/\sigma$. And so on.

Theorem 6 is proved. □

5. Languages with weak linear and strong nonlinear structure

Recall that a grammar is called *unambiguous* if every generated word has a single derivation tree.

Definition 8. We shall say that a language has a *weak linear* (*strong nonlinear*) structure if it can be generated by an unambiguous weak linear (correspondingly strong nonlinear) grammar.

Theorem 9. *The family of all unambiguous grammars generating languages with weak linear structure and the family of all unambiguous grammars generating languages with strong nonlinear structure are algorithmically separable.*

Proof. The following decidable property of a grammar separates these two families: "the domain of convergence of grammar's FPS is open". We must prove that this property separates the families and is decidable.

Let G_1 be an unambiguous grammar generating a language L with weak linear structure. As G_1 is unambiguous, its power series is $\sum_{w \in L} w \cdot 1$. On the other hand, there is an unambiguous weak linear grammar G_2 generating L. The formal power series of G_2 is also $\sum_{w \in L} w \cdot 1$. By Theorem 6 the domain of convergence of this series is open.

Let G_1 be an unambiguous grammar generating a language L with strong nonlinear structure. Let us prove that the domain of convergence of its power series is not open. Its power series $\sum_{w \in L} w \cdot 1$ is in the same time the power series of an unambiguous strong nonlinear grammar G_2. Therefore, the domain of divergence of $\sum_{w \in L} w \cdot 1$ is open. The range of the variables of series is a connected region, consequently if the domain of convergence of $\sum_{w \in L} w \cdot 1$ is also open, then one of these two domains is empty.

The domain of convergence includes a neighborhood of the point $(0, \dots, 0)$. Indeed, an unambiguous series has not more than m^l terms of length l (where m stands for the number of variables). Hence the series converges if all variables are less than $1/m$ because it is less than a geometric progression.

The domain of divergence contains the point $(1,\dots,1)$. This follows from the infiniteness of L.

Let us prove now that given a grammar we can effectively decide whether the domain of convergence of its power series is open.

Let us consider system (2) for this grammar. Let $c_1,\dots,c_m$ be positive reals. Let r_i stands for the right-hand side of the ith equation in (2), that is, $r_i = u_1^i + \cdots + u_{k_i}^i$. As r_i has a finite number of terms, after substituting $c_1,\dots,c_m$ for $t_1,\dots,t_m$ we get a power series with a finite number of terms in variables $x_1,\dots,x_n$. Let us denote this series by $\bar{r}_i[x_1,\dots,x_n]$. We shall regard this series as a polynomial of (commuting) variables $x_1,\dots,x_n$. For every power series s of variables $t_1,\dots,t_m$ with nonnegative coefficients by $\bar{s}$ we denote the numerical series obtained by substituting $c_1,\dots,c_m$ for $t_1,\dots,t_m$ in s (the nonnegativeness of coefficients provides that the convergence does not depend on the order of terms).

Lemma 10. *Let $\langle s_1,\dots,s_n\rangle$ be a solution of system* (2). *Then the numerical series $\bar{s}_1$ converges iff the algebraic system*

$$x_i = \bar{r}_i[x_1,\dots,x_n], \quad i = 1,\dots,n \tag{7}$$

has a positive real solution.

Proof. Assume that the series $\bar{s}_1$ converges. We claim that in this case for all i the series $\bar{s}_i$ converges. In fact, this claim was proved in the proof of Theorem 6, part II. Let s_i^+ be the sum of the series $\bar{s}_i$. Then $(s_1^+ \dots s_n^+)$ is the positive solution of system (2) (and hence (7)), because the rules of addition and multiplication of FPS are well defined for converging series.

Conversely, let $\bar{s}_1$ diverge. Let us prove that system (7) has no positive solution. Assume the contrary: $(a_1 \dots a_n)$ is a positive solution of system (7), that is, $a_i = \bar{r}_i[a_1,\dots,a_n]$, $i = 1,\dots,n$. Let p_i^α be the power series defined in the proof of Theorem 2. Let $\bar{p}_i^\alpha$ be the result of substituting $c_1,\dots,c_m$ for $t_1,\dots,t_m$ in p_i^α. Clearly, $\bar{p}_i^\alpha$ are polynomials in $x_1,\dots,x_n$ satisfying the equalities $\bar{p}_i^0 = x_i$, $\bar{p}_i^{\alpha+1} = \bar{r}_i[\bar{p}_1^\alpha,\dots,\bar{p}_n^\alpha]$. Therefore, $a_i = \bar{p}_i^\alpha[a_1,\dots,a_n]$ for all $\alpha \in \mathbb{N}$ and all i (this can be proved by induction). Consequently for all α we have $a_1 = \bar{p}_1^\alpha[a_1,\dots,a_n] \geqslant \bar{p}_1^\alpha[0,\dots,0] = \bar{s}_1^\alpha$ (where s_i^α were defined in the proof of Theorem 2). We claim that $\lim_{\alpha\to\infty} \bar{s}_1^\alpha = +\infty$. Indeed, for each $w \in \{t_1,\dots,t_m\}^*$

$$\lim_{\alpha\to\infty} (s_1^\alpha, w) = (s_1, w)$$

and the numerical series $\bar{s}_1$ diverges, hence $\lim_{\alpha\to\infty} \bar{s}_1^\alpha = +\infty$. Thus we have got a contradiction: $a_1 \geqslant +\infty$.

Lemma 10 is proved. □

Now, following [6], in order to construct an algorithm deciding, whether the domain of convergence of grammar's FPS is open, we use the fact that the elementary theory of real field is decidable. Due to Lemma 10 for each grammar we can construct a formula $\varphi(c_1 \dots c_m)$ of this theory such that the FPS of the grammar converges in a point $(c_1 \dots c_m)$ iff $\varphi(c_1 \dots c_m)$ is true. Using $\varphi(c_1 \dots c_m)$, we can construct a formula ψ that is true iff the domain of convergence of FPS is open.

Theorem 9 is proved. □

6. Noninvariance

A property P of an unambiguous grammar is called *invariant* if for any language generated by an unambiguous grammar satisfying P any grammar that generates this language also satisfies P. In connection with above results the following question arises: are the properties of unambiguous grammar to be weak linear and to be strong nonlinear invariant? The answer is unfortunately negative. The language of all nonempty words in a binary alphabet (obviously having weak linear structure) can be generated by some unambiguous grammar, which is not weak linear.

Theorem 11. *Consider the grammar* $\{F \to S,\ F \to T,\ F \to TF,\ S \to f,\ S \to fS,\ S \to fT,\ S \to fTS,\ T \to l,\ T \to fTT\}$, *where* F, S, T *are nonterminals*, F *is the initial nonterminal*, f, l *are terminals. This grammar is unambiguous*, *is not weak linear*, *and generates the language of all nonempty words over the alphabet* $\{f, l\}$.

Remark. For brevity the grammar has rules of the form nonterminal → nonterminal. However the canonical transformation of the grammar to the reduced form preserves all properties listed in the theorem.

Proof. Obviously, the grammar is not weak linear.

Let us explain the sense of the grammar. The nonterminals mean: F—“forest”, T—“tree”, S—“shoot”; terminals mean: f—“fork”, l—“leaf”.

Definition 12. A word w over the alphabet $\{f, l\}$ is called *a tree* (*a shoot*) if w can by derived from the nonterminal T (correspondingly S).

The geometrical image of the tree l is a single point. If u and v are trees, then the geometrical image of the tree fuv consists of the root and the geometrical images of u and v growing from its two sons.

Let us describe the analysis of any word according to our grammar. We try to find a beginning u of the word such that u is a tree. If we are successful, then we erase u and try again to find a beginning that is a tree. And so on until the remaining word becomes the empty word or a shoot.

The unambiguousness of the grammar follows from two lemmas.

Lemma 13. *If a tree is a prefix of another tree* (*in this case we say that the trees are consistent*), *then these trees are equal.*

Proof. By induction on the sum of trees’ lengths. Let u and v be consistent trees. If one of them has length 1, then the assertion is evident. Else $u = fu_1u_2$ and $v = fv_1v_2$, where u_1, u_2, v_1, and v_2 are trees. We see that v_1 and u_1 are consistent. By induction hypothesis $u_1 = v_1$. Therefore, u_2 and v_2 are consistent. By induction hypothesis $u_2 = v_2$. Hence $u = v$. Lemma 13 is proved. □

Lemma 14. *No tree is a prefix (even improper) of a shoot.*

Proof. By induction on shoot's length. Suppose $w = uv$, where w is a shoot and u is a tree. Then the word u has the prefix f. Consider two cases.

Case 1: $w = f$. Obviously we get a contradiction.

Case 2: $w = fz$, z is nonempty. As $fz = uv$, the tree u has the form fxy, where x and y are trees. Therefore, $z = xyv$.

As fz is a shoot, three cases are possible: z is a shoot, z is a tree and z is the concatenation of a tree and a shoot.

The case "z is a shoot" contradicts the induction hypothesis.

If z is a tree, then $z = x$ by Lemma 13 and therefore the tree y is empty; this is impossible.

If $z = z_t z_s$, where z_t is a tree and z_s is a shoot, then $z_t z_s = xyv$. Therefore, $z_t = x$ by Lemma 13. Thus $z_s = yv$; this contradicts the induction hypothesis.

Lemma 14 is proved. □

The following lemma shows that each word can be derived from the initial nonterminal.

Lemma 15. *If no prefix of a nonempty word is a tree, then this word is a shoot.*

Proof. By induction on the length of the given nonempty word u.

If the first letter of u is l, then u begins with a tree.

If $u = f$, then u is a shoot.

If $u = fv$, where v is nonempty and has no tree prefix, then by the induction hypothesis v is a shoot; therefore u is also a shoot.

Let $u = fvw$ and v be a tree. If w is a shoot, then u is also a shoot. If w is empty, then u is again a shoot. Else by the induction hypothesis w has a tree prefix (say) z; therefore u has a tree prefix fvz.

Lemma 15 is proved. □

As to the unambiguousness of this grammar, we can use Lemmas 13–15 to prove by induction on length of a word u that from each nonterminal the word u has not more than one derivation tree.

Theorem 11 is proved. □

Let us prove that the property of an unambiguous grammar to be strong nonlinear is also noninvariant. We shall say that any word over $\{f, l\}$ derived from the nonterminal T in the grammar of Theorem 11 is a *tree*. Evidently, the language of all trees is a language with strong nonlinear structure.

Theorem 16. *Consider the grammar* $\{T_0 \to l,\ T_0 \to fT_0T_0,\ T_1 \to l,\ T_1 \to flT_1,\ T_1 \to ffT_0T_0T_0\}$, *where* T_0, T_1 *are nonterminals*, T_1 *is the initial nonterminal*, f, l *are* *terminals. This grammar is unambiguous, is not strong nonlinear, and generates the set of all trees.*

Proof. Obviously, the nonterminal T_1 is not nonlinear and the nonterminal T_0 is nonlinear. The sequence $T_1, T_1, T_1, \dots$ is the infinite quasiderivation, which does not contain the nonlinear terminal (T_0). Therefore this grammar is not strong nonlinear. Obviously, the words derived from T_0 are exactly the trees. We can also prove by induction that only trees can be derived from T_1.

Let us prove that every tree has a unique derivation from T_1 by induction (the parameter of induction is the length of the tree u).

Evidently, if $u = l$, then u has a unique derivation. Else $u = fu_1u_2$, where u_1 and u_2 are trees.

If $u_1 = l$, then any derivation of u must begin with the rule $T_1 \to flT_1$. By the induction hypothesis u_2 has a unique derivation from T_1.

If $u_1 \neq l$, then $u_1 = fv_1v_2$, where v_1 and v_2 are trees. In this case any derivation of u must begin with the rule $T_1 \to ffT_0T_0T_0$. By Theorem 11 every tree has a unique derivation from T_0.

Theorem 16 is proved. □

7. Conclusion

The results of [6] and our Theorem 9 show that the family of unambiguous grammars has some "good" algorithmic properties, which the family of all context-free grammars does not have. But significance of this fact is partially reduced because the property of a grammar "to be unambiguous" is undecidable. Nevertheless the state of affairs might be considered good if there would be a recursively enumerable family of unambiguous grammars such that its members generate all unambiguous languages. The following theorem proved by Gorbunov shows that this is not the case.

Theorem 17 (see Gorbunov [1,2]). *There is no recursively enumerable set of context-free grammars (not necessary unambiguous) such that the grammars of this set generate exactly all unambiguous languages.*

The question whether such a family does exist is an example of a problem of a certain type which, seems to be interesting. Let we have an enumeration of some family of objects and we have a set M of numbers with good algorithmic properties. Is there a recursively enumerable subset of M such that each object that has a number in M has also a number in this subset?

The author thinks that the problem "to enumerate a subset of M such that each object that has a number in M has also a number in this subset" is more actual than the traditional problem of decision of M.

Acknowledgements

Author is grateful to N.K. Vereshchagin for his help in writing the draft version of the present paper.

References

[1] K.Yu. Gorbunov, On one algorithmic problem in mathematical linguistics, Proc. Symp. on Semiotical Aspects of Formalization of Intellectual Activity held in Borzhomi (Georgia), Moscow, 1988, pp. 8–11 (in Russian).

[2] K.Yu. Gorbunov, There is no recursively enumerable set of context free grammars generating the class of unambiguous languages, Mat. Zametki 50 (1) (1991) 34–40 (in Russian).

[3] An.A. Muchnik, An application of Semenov's method to analysis of structure of context-free languages, Proc. Symp. on Semiotical Aspects of Formalization of Intellectual Activity held in Kutaisi (Georgia), Moscow, 1985, pp. 212–214 (in Russian).

[4] An.A. Muchnik, One application of real-valued interpretation of grammatical series, preprint, Institute of New Technologies, Moscow (in Russian).

[5] A. Salomaa, M. Soittola, Automata-Theoretic Aspects of Formal Power Series, Springer, New York, 1978.

[6] A.L. Semenov, Algorithmic problems for power series and context-free grammars, Dokl. Akad. Nauk SSSR 212 (1973) 50–52 (in Russian).