Archivum Mathematicum

Milan Drášil

A grammatical inference for C-finite languages

Archivum Mathematicum, Vol. 25 (1989), No. 3, 163--173

Persistent URL: http://dml.cz/dmlcz/107353

ARCHIVUM MATHEMATICUM (BRNO)
Vol. 25, No. 3 (1989), 163–174

A GRAMMATICAL INFERENCE FOR C-FINITE LANGUAGES

MILAN DRÁŠIL

(Received September 15, 1986)

Abstract. For any language L, any finite set of contexts C, and any positive integer i we construct a linear grammar $FG(L, C, i)$ generating a language, whose ith fragment coincides with the ith fragment of the given language. If there exists some positive integer k such that for any $i \geq k$ the grammars $FG(L, C, i)$ and $FG(L, C, k)$ coincide, then the grammar $FG(L, C, k)$ generates the given language. A necessary and sufficient condition for this coincidence is given.

Key words. Grammatical inference, linear grammar, context, derivative, C-finite language complete set of contexts.

MS Classification. 68 Q 50.

1. INTRODUCTION

In special cases of grammars (e.g. regular, linear or context-free ones) nonterminal symbols can be considered the sets of all words generated by them. M. Novotný and his collaborators investigate possibilities of constructing grammars, where the role of nonterminals is played by special sets of words, so called derivatives and syntactic categories. The noneffective constructions based on this idea can be seen in [1], [7], [8], [10], the effective ones in [6], [11]. Similar ideas are used in algorithm inferring a linear harmonic grammar, which has been proposed by K. Tanatsugu [12].

This paper presents an effective algorithm inferring a linear grammar from a sample called fragment of the language (the set of all words of the language that are not larger than a given positive integer). The idea of using derivatives as nonterminals in effective constructions is due to M. Novotný ([9]).

2. PRELIMINARY DEFINITIONS AND NOTATION

By N we denote the set of all positive integers. An *alphabet* V is a finite set, whose elements are called *symbols*. The set of all words over an alphabet V –

including the empty word λ – is denoted by V^*. For any $x, y \in V^*$ we denote by xy their concatenation and for any $P, Q \subseteq V^*$ we put $PQ = \{xy;\ x \in P,\ y \in Q\}$. For any $a \in V$ a^k denotes the word of k concatenated a's. The lenght of the word x denoted by $|x|$ is the number of symbols used in its formation. An element $(u, v) \in V^* \times V^*$ is called a *context over* V or simply a *context*. We put $|(u, v)| = |u| + |v|$. For two arbitrary contexts $w_1 = (u_1, v_1)$ and $w_2 = (u_2, v_2)$ we define the operation $w_1 \circ w_2 = (u_1 u_2, v_2 v_1)$ and it is easy to see that $(V^* \times V^*, \circ, (\lambda, \lambda))$ is a monoid. Any set of contexts C generates the submonoid in the above mentioned monoid. By $[C]$ we denote its carrier (i.e. any $w \in [C]$ is of the form $w = w_1 \circ \ldots \circ w_k$, where $w_1, \ldots, w_k \in C$). A *language* L over an alphabet V is an arbitrary subset of V^*. For any $Q \subseteq V^*$ we put $\|Q\| = \max\{|t|;\ t \in Q\}$ if Q is finite $\|Q\| = \infty$ otherwise. A *grammar* is an ordered quadruple $G =$ fragment of the set Q. Let $w = (u, v)$ be a context and $Q \subseteq V^*$. Then the set Q_w $= (S, V, R, s_0)$, where V and S are disjoint alphabets called *terminal* and *nonterminal* ones respectively, $R \subseteq (V \cup S)^* \times (V \cup S)^*$ finite set of *rules* and $s_0 \in S$ starting symbol. The relation of *direct derivation* denoted by $\rightarrow$ and its transitive-reflexive closure denoted by $\rightarrow^*$ are defined in the usual manner. Grammars are said to be *regular* and *linear*, if their sets of rules are of the form $R \subseteq S \times V^* \cup S \times V^* S$ and $R \subseteq S \times V^* \cup S \times V^* S V^*$ respectively. We put $L(G) = \{t;\ t \in V^*,\ s_0 \rightarrow^* t\}$ and $L(G)$ is said to be the *language generated by grammar* G. For any positive integer i and any $Q \subseteq V^*$ the set $iQ = \{t;\ t \in Q,\ |t| \leq i\}$ is called *ith fragment of the set* Q. Let $w = (u, v)$ be a context and $Q \subseteq V^*$. Then the set $Q_w = \{t;\ utv \in Q\}$ is said to be the derivative of the set Q by the context w. Clearly $(Q_x)_y = Q_{x \circ y}$ for any contexts $x, y \in V^* \times V^*$ and any set $Q \subseteq V^*$. For any sets P, $Q \subseteq V^*$ we set $P \Subset Q$ if and only if there exists some positive integer i such that P is the ith fragment of Q. Obviously for any system of sets $T \subseteq 2^{V^*}$ the pair $(T, \Subset)$ is a partially ordered set.

3. CONSTRUCTION OF FG-GRAMMARS

Let L be an arbitrary language over an alphabet V, C finite set of nontrivial contexts (i.e. contexts different from (λ, λ)). We set

$$P(i) = \{(iL)_w;\ w \in [C],\ (iL)_w = \emptyset\} \cup \{iL\}.$$

(Many constructions in this paper depend on fixed sets L and C. For the sake of notation convenience we shall omit them as parameters.)

Clearly $(iL)_w = \emptyset$ for any $w \in [C]$ with the property $|w| > i$, thus the set $P(i)$ is finite. By $M(i)$ we denote the set of all maximal elements in the ordered set $(P(i), \subset)$. Note that $iL \in M(i)$. Let us have a mapping $^-$ of $\bigcup_{i \in N} P(i)$ into $\bigcup_{i \in N} M(i)$ with the following properties:

(i) $Q \in P(i)$ implies $\bar{Q} \in M(i)$,
(ii) $Q \subseteq \bar{Q}$.

Any mapping with those properties will be called a *C-mapping of the language L.* Any pair $(Q, u\bar{Q}_w v)$, where $Q \in M(i)$, $w \in C$ and $Q_w \in P(i)$ is said to be an *FG-rule of the ith fragment*. Now, let us define the mapping c of $\bigcup_{i \in N} \{\{i\} \times P(i)\}$ into N in the following way:

$$c(i, Q) = \max \{i - |w|;\ w \in [C],\ Q = (iL)_w\}.$$

An arbitrary *FG*-rule of the ith fragment (Q, uPv) is said to be *suitable* if for any $t \in \{u\}\,P\{v\} - Q$ the condition $|t| > c(i, Q)$ holds. Now we can construct the grammar $FG(L, C, i)$ belonging to the ith fragment of the language L. We put

$R_1(i)$ – the set of all suitable *FG*-rules of the ith fragment,

$R_2(i) = \{(Q, t);\ Q \in M(i),\ t \in Q - \{urv;\ (Q, uPv) \in R_1(i),\ r \in P\}\}$.

The ordered quadruple $FG(L, C, i) = (V, M(i),\ R_1(i) \cup R_2(i),\ iL)$ is a linear grammar, where we suppose without loss of generality that the sets V and $M(i)$ are disjoint. In the next section we show that the construction of a grammar $FG(L, C, i)$ is relatively independent on mapping $^-$, the only importance is that it has the properties of a C-mapping.

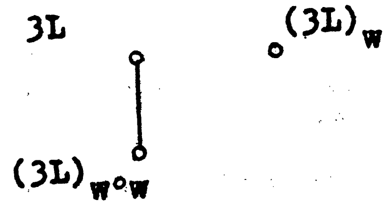

fig. 1

3.1 Example. (a) Let $V = \{a\}$, $C = \{w = (a, \lambda)\}$ and $3L = \{\lambda, a^2\}$. The ordered set $(P(3), \subset)$ is shown in fig. 1. We have two *FG*-rules $3L \to a(3L)_w$ and $(3L)_w \to a3L$ and it is easy to see that both ones are suitable. Thus the grammar $FG(L, C, 3)$ contains the following rules:

$3L \to a(3L)_w \mid \lambda,$
$(3L)_w \to a3L.$

This grammar generates all even powers of the symbol a. (b) Let V and C be the same ones as in (a) but assume that the sample $\{\lambda, a^2\}$ is the fourth fragment of some language. The ordered set $(P(4), \Subset)$ is of the same structure as in (a) but the rule $(4L)_w \to a4L$ is not suitable since $a^3 \in \{a\}\, 4L - (4L)_w$ and $c(4, (4L)_w) = 3$. Thus we obtain the grammar $FG(L, C, 4)$ with the rules:

$$4L \to a(4L)_w \mid \lambda,$$
$$(4L)_w \to a.$$

This grammar generates exactly the given sample. (c) If the fourth fragment of some language $4L = \{\lambda, a^2, a^4\}$ and $C = \{(a, \lambda)\}$, the construction of the grammar $FG(L, C, 4)$ leads to the same one as in (a) (up to renaming nonterminals). □

The example 3.1. shows that the grammar $FG(L, C, i)$ generates all words of iL. Moreover the suitability of FG-rules guarantees that if the grammar $FG(C, L, i)$ generates some words that are not contained in iL, then they must be larger than i. Let us prove this fact exactly. In what follows we suppose that we are given fixed sets V, L and C. □

3.2. Lemma *Let $i \in N$, $Q \in P(i)$ and $w \in C$ such that $Q_w \in P(i)$. Then:*

(i) *$t \in Q$ implies $|t| \leq c(i, Q)$;*
(ii) *$c(i, iL) = i$,*
(iii) *$c(i, Q) - |w| \leq c(i, Q_w)$.*

Proof. The statements (i) and (ii) are trivial, we prove (iii). Let $x \in [C]$ be a context such that $Q = (iL)_x$ and $c(i, Q) = i - |x|$. We have $c(i, Q_w) = c(i, (iL)_{x\circ w}) = \max\{i - |y|;\, y \in [C], (iL)_{x\circ w} = (iL)_y\} > i - |x \circ w| = i - |x| - |w| = c(i, Q) - |w|$. □

3.3. Lemma *For any $i \in N$ the following assertions hold.*

(i) *$Q \in M(i)$ and $t \in Q$ imply $Q \to^* t$ in the grammar $FG(L, C, i)$,*
(ii) *$L(FG(L, C, i)) \supseteq iL$.*

Proof. (i) By induction on lenght of the word t.

(a) If $|t| = 0$ (i.e. $t = \lambda$), then there exists the rule $Q \to \lambda$ in $R_2(i)$ since $\lambda \notin \{u\}\, P\{v\}$ for any rule $Q \to uPv$ in $R_1(i)$.

(b) Let $|t| > 0$ and suppose that the assertion holds for any word r such that $|r| < |t|$. If $R_1(i)$ does not contain any rule $Q \to uPv$ with the property $t = urv$, then $R_2(i)$ contains the rule $Q \to t$. If $R_1(i)$ contains some rule $Q \to uPv$ such that $t = urv$, then $P = Q_w$ where $w = (u, v)$, $r \in Q_w$ and $r \in P$ since Q_w is a fragment of P. Furthermore $|r| < |t|$ implies $P \to^* r$ and $Q \to uPv \to^* urv = t$ completes the proof of the assertion (i).

(ii) is a consequence of (i). □

3.4. Lemma *For any suitable FG-rule of the ith fragment $Q \to uPv$ holds*:
$c(i, Q) - |(u, v)| \leq c(i, P)$.

Proof. Let $(u, v) = w$. If $P = Q_w$, then by 3.2. (iii) the assertion holds. Assume that $P = \bar{Q}_w \neq Q_w$. Then there exists some word t with the property $t \in P$ and $t \notin Q_w$ since Q_w is a fragment of P. Consequently $utv \in \{u\} P\{v\} - Q$ and this implies $|utv| > c(i, Q)$ since the rule $Q \to uPv$ is suitable. By 3.2. (i) we have $|t| \leq c(i, P)$ and $c(i, Q) - |w| < |t| \leq c(i, P)$ completes the proof. □

3.5. Lemma *For any $i \in N$ the following assertions hold.*
(i) *$Q \to^* t$ in the grammar $FG(L, C, i)$ and $|t| \leq c(i, Q)$ imply $t \in Q$.*
(ii) *$iL \supseteq iL(FG(L, C, i))$.*

Proof. (i) By induction on lenght of derivation.
(a) If t can be derived in one step from Q, then there exists a rule $Q \to t$ in $R_2(i)$ and $t \in Q$ trivially.
(b) Suppose that t can be derived in n steps ($n > 1$) and that the assertion holds for any $k < n$. Consequently there exists a rule $Q \to uPv$ such that $t = urv$ and r can be derived from P in $n - 1$ steps. We have $|t| = |urv| \leq c(i, Q)$, i.e. $|r| \leq \leq c(i, Q) - |(u, v)|$ and by 3.4. $c(i, Q) - |(u, v)| \leq c(i, P)$. Thus $|r| \leq c(i, P)$ and $r \in P$. Finally $t = urv \in \{u\} P\{v\}$ and $|t| \leq c(i, Q)$ implies $t \in Q$ since otherwise we would have a contradiction with the suitability of the FG-rule $Q \to uPv$.
(ii) is a consequence of (i) and 3.2. (ii). □
3.3. (ii) and 3.5. (ii) yield the following result.

3.6. Theorem $iL = iL(FG(L, C, i))$. □
A language L is said to be *FG-grammatizable*, if there exists a finite set of nontrivial contexts C, C-mapping$^-$ and a positive integer k such that for any $i \geq k$ the grammars $FG(L, C, i)$ and $FG(L, C, k)$ coincide up to renaming nonterminals. □

4. C-FINITE LANGUAGES, COMPLETE SETS OF CONTEXTS

Let L be an arbitrary language over an alphabet V, C a finite set of nontrivial contexts. We define the equivalence relation R on $[C]$ in the following way:
For any $x, y \in [C]$ xRy if and only if $L_x = L_y$. A language L is said to be C-finite if the set $[C]/R$ is finite (c.f. [10]).

4.1. Lemma *$(iL)_x \subseteq (iL)_y$ holds for any $i \in N$ and any $x, y \in [C]$ such that xRy and $|y| \leq |x|$.*
Proof. If $|y| > i$ or $|x| > i \geq |y|$, then the assertion is trivial. Let $x = (x_1, x_2)$, $y = (y_1, y_2)$ and assume that $i \geq |x| \geq |y|$. First we prove $(iL)_x \subseteq$

$\subseteq (iL)_y$. For any $t \in (iL)_x$ we have $x_1 t x_2 \in iL$, consequently $x_1 t x_2 \in L$ and xRy implies $y_1 t y_2 \in L$. Furthermore $|y_1 t y_2| \leq |x_1 t x_2| \leq i$, hence $y_1 t y_2 \in iL$ and $t \in (iL)_y$. Now we prove that for any $t \in (iL)_y$ with the property $t \leq \max\{|r|; r \in (iL)_x\} = m$ the condition $t \in (iL)_x$ holds. Let $t \in (iL)_y$ and $|t| \leq m$. Similarly we have $x_1 t x_2 \in L$ and clearly $m \leq i - |x|$. Thus $|x_1 t x_2| = |t + |x| \leq m + i - - m = i$ and consequently $t \in (iL)_x$ which completes the proof. □

4.2. Lemma *Let $x, y \in [C]$ be two contexts such that L_x is infinite and $x\bar{R}y$. Then there exists $k \in N$ such that for any $i \geqq k$ $(iL)_x$ is not a fragment of $(iL)_y$.*

Proof. Let $x = (x_1, x_2)$ and $y = (y_1, y_2)$. $x\bar{R}y$ implies that there exists a word $t \in (L_x - L_y) \cup (L_y - L_x)$. If there exists $t \in L_x - L_y$, then $x_1 t x_2 \in L$ and $y_1 t y_2 \notin L$. We put $k = |x_1 t x_2|$. Obviously $t \in (iL)_x - (iL)_y$ for any $i \geqq k$, hence $(iL)_x$ is not a subset of $(lL)_y$. The second subcase $t \in L_y - L_x$ implies that $y_1 t y_2 \in L$ and $x_1 t x_2 \notin \notin L$. We put $k \geqq |t|$ sufficiently large such that there exists $u \in (kL)_x$ with the property $|u| \geqq |t|$ (this is possible since L_x is an infinite set of words). For any $i \geqq k$ we have $t \in (iL)_y - (iL)_x$ and $|t| \leq \max\{|u|; u \in (iL)_x\}$. Thus $(iL)_x$ is not a fragment of $(iL)_y$. □

4.3. Lemma *Let $x, y \in [C]$ be two contexts such that L_x is a finite set. Then there exists $k \in N$ such that for any $i \geqq k$ $(iL)_x \Subset (iL)_y$ if and only if $(kL)_x \Subset (kL)_y$.*

Proof. Let $x = (x_1, x_2)$, $y = (y_1, y_2)$ and let us set $k = \max\{|u|; u \in L_x\} + + \max\{|x|, |y|\}$. Clearly $(iL)_x = L_x$ for any $i \geqq k$.

(a) We prove „if" part of the assertion. Let $(kL)_x \Subset (kL)_y$ and $i \geqq k$. Obviously $(iL)_x = (kL)_x \subseteq (kL)_y \subseteq (iL)_y$. Assume that there exists a word $t \in (iL)_y - (iL)_x$ (otherwise $(iL)_x = (iL)_y$ and „if" part of the proof is trivial). If $t \in (kL)_y$ then $|t| > > \max\{|u|; u \in (iL)_x\}$ since $(iL)_x = (kL)_x \Subset (kL)_y$. If $t \notin (kL)_y$, then $|y_1 t y_2| > k$, i.e. $|t| > k - |y| = \max\{|u|; u \in (iL)_x\} + \max\{|x|, |y|\} - |y| \geqq \max\{|u|; u \in (iL)_x\}$.

(b) To prove „only if" part of the assertion let us suppose that $(kL)_x$ is not a fragment of $(kL)_y$. If there exists $t \in (kL)_x - (kL)_y$, then $t \in (iL)_x - (iL)_y$ for any $i \geqq k$ since otherwise $t \in (iL)_y$ implies $|y_1 t y_2| > k$, i.e. $|t| > k - |y| \geqq \geqq \max\{|u|; u \in (iL)_x\}$ which would be a contradiction. If there exists $t \in (kL)_y - - (kL)_x$ with the property $|t| \leq \max\{|u|; u \in L_x\}$, then clearly $t \in (iL)_x = (kL)_x$ and consequently $t \in (iL)_y - (iL)_x$ for any $i \geqq k$. □

4.4. Lemma *Let the set $\{L_w; w \in [C], L_w$ is infite$\}$ be finite. Then the set $\{L_w; w \in [C], L_w$ is finite$\}$ is finite too.*

Proof. If L is finite the assertion is trivial, suppose that L is infinite. Let $n \geqq 1$ be an integer such that for any infinite derivative Q of L by the context from $[C] - - \{(\lambda, \lambda)\}$ there exist contexts $w_1, \ldots, w_k \in C$ such that $k < n$ and $Q = L_w$ where $w = w_1 \circ \ldots \circ w_k$. Setting $m = \max\{\{0\} \cup \{\|L_w\|; L_w$ is finite, $w = w_1 \circ \ldots \circ w_k$,

$w_i \in C$ for $1 \leq i \leq k \leq n\}\}$ we prove that for any finite derivative Q the condition $\| Q \| \leq m$ holds. Let L_w be an arbitrary finite derivative where $w = w_1 \circ \ldots \ldots \circ w_s$ and $w_i \in C$ for $1 \leq i \leq s$. If L_{w_1} is finite, then clearly $\| L_w \| \leq \| L_{w_1} \| \leq m$. Assume that $\| L_{w_1} \|$ is infinite and let j be an integer with the following property; setting $x = w_1 \circ \ldots \circ w_j$ L_x is infinite and $L_{x \circ w_{j+1}}$ is finite. There exists a context $y = y_1 \circ \ldots \circ y_k$ where $y_i \in C$ for $1 \leq i \leq k$, $k < n$ and $L_x = L_y$. We have $\| L_{x \circ w_{j+1}} \| = \| L_{y \circ w_{j+1}} \| \leq m$ and clearly $\| L_w \| \leq \| L_{x \circ w_{j+1}} \|$. □

4.5. Corollary *For any language L and any finite set of nontrivial contexts C the following statements are equivalent*:

(i) *L is C-finite.*

(ii) *There exists $m \in N$ such that* card $(M(i)) \leq m$ *for any $i \in N$.*

Proof. By 4.1. (i) implies (ii) since it sufices to put $m = \text{card}([C]/R)$. Conversely suppose that L is not C-finite. We set $D = \{L_w;\, w \in [C],\, L_w \text{ is infinite}\}$ and by 4.4. D is an infinite set. Furthermore by 4.1. and 4.2. for any two different derivatives P, $Q \in D$ there exist contexts $x, y \in [C]$ and an integer k such that $P = L_x$, $Q = L_y$ and for any $i \geqq k (iL)_x \neq (iL)_y$ and $(iL)_x, (iL)_y \in M(i)$. This completes the proof. □

In what follows we show that for any language L and any finite set of nontrivial contexts C there exists an integer k and *a* C-mapping $\bar{}$ such that for any $i \geqq k$ the sets of FG-rules $R_1(i)$ and $R_1(k)$ coincide if and only if L is C-finite. The necessity of this condition follows by 4.5., we show sufficiency. Let L be a C-finite language and $D = \{Q_1, \cdots, Q_n\}$ be the set of all derivatives of L by the contexts from $[C]$. We choose the set of contexts $Y = \{y_1, \cdots, y_n\} \subseteq [C]$ in the following way:

(i) $Q_i = L_{y_i}$ for $1 \leq i \leq n$,

(ii) $x \in [C]$ and xRy_i imply $| y_i | \leq | x |$.

4.1. guarantees $M(i) \subseteq \{(iL)_y;\, y \in Y\}$. Let us put $C_0 = C \cup \{(\lambda, \lambda)\}$. By 4.1., 4.2., 4.3. and construction of Y it follows that for any contexts $x, y \in Y$ and $w \in C_0$ there exists an integer k_{pwy} such that for any $i \geq k_{xwy}$ $(iL)_{x \circ w} \Subset (iL)_y$ if and only if $(k_{xwy}L)_{x \circ w} \Subset (k_{xwy}L)_y$. We put $k = \max\{k_{xwy};\, x, y \in Y,\, w \in C_0\}$. We have $(iL)_{x \circ w} \Subset (iL)_y$ if and only if $(kL)_{x \circ w} \Subset (kL)_y$ for any x, $y \in Y$, $w \in C_0$. Furthermore if $(iL)_x = (iL)_y$ for some $x, y \in Y$ and $i \geqq k$, then $x = y$ since by construction of the index k $(iL)_x = (iL)_y$ holds for any $i \geqq k$, i.e. $L_x = L_y$. Denoting by X the subset of Y such that $M(k) = \{(kL)_x;\, x \in X\}$ we can estabilish the following assertion.

4.6. Lemma *Let L be a C-finite language. Then there exists $k \in N$ and a finite set of contexts $X \subseteq [C]$ such that for any $i \geqq k$ hold*:

(i) $M(i) = \{(iL)_x;\, x \in X\}$, *$x, y \in X$ and $x \neq y$ imply $(iL)_x \neq (iL)_y$.*

(ii) $(iL)_{xow} \subseteq (iL)_y$ *if and only if* $(kL)_{xow} \subseteq (kL)_y$ *for any* $x, y \in X$ *and any* $w \in [C]$.

(iii) $c(i, (iL)_x) = i - |x|$.

Proof. Let Y, $X \subseteq Y$ and k be the above constructed sets and index. (i) and (ii) has been already proved, we prove (iii). Assume that $c(i, (iL)_x) > i - |x|$ for some $x \in X$ and $i \geq k$. Consequently there exists a context $w \in [C]$ such that $(iL)_x = (iL)_w$ and $|w| < |x|$, by construction of the set X we have $x\bar{R}w$. Let $z \in Y$ and $y \in X$ be the contexts such that wRz and $(iL)_z \subseteq (iL)_y$. We have $(iL)_x \subseteq (iL)_W = \subseteq (iL)_z \subseteq (iL)_y$ and this implies $(iL)_x = (iL)_y$ since $(iL)_x, (iL)_y \in M(i)$. Thus $(iL)_x = (iL)_z$, consequently $x = z$ and we have $x\bar{R}wRz = x$ which is a contradiction. □

Let L be a C-finite language, X a set of contexts and let k be the least integer such for any $i \geq k$ the conditions 4.6. (i), (ii) and (iii) hold. Then X is said to be the *principal set* of *contexts of the language* L and $k = d_1(L, C)$ is said to be the first degree of the language L.

4.7. Lemma *Let* L *be a* C*-finite language,* X *its principal set of contexts and let* $\tilde{}$: $P(d_1(L, C)) \to M(d_1(L, C))$ *be an arbitrary mapping with the property* $Q \subseteq \tilde{Q}$. *Then there exists a* C*-mapping* $\bar{}$ *such that hold*:

(i) $\bar{Q} = \tilde{Q}$ *for any* $Q \in P(d_1(L, C))$.

(ii) *The sets of FG-rules of the* $d_1(L, C)$th *and ith fragment coincide for any* $i \geq d_1(L, C)$.

Proof. By 4.6. (ii) it suffices to put $\overline{(iL)_{xow}} = (iL)_y$ if and only if $(\widetilde{kL})_{xow} = (kL)_y$ for any $i \geq k = d_1(L, C)$, any $x, y \in X$ and any $w \in C$. □

4.7. guarantees not only the existence of the C-mapping $\bar{}$ but also the independence of choice of restriction $\bar{}$ on $P(i)$ for any $i \in N$. In other words we can construct the restriction $\bar{}$ on $P(i)$ arbitrarily, i.e. effectively. Any mapping with the property 4.7. (ii) will be called *a principal mapping* of the language L. Now we can estabilish the assertion guaranteeing coincidence of the sets $R_1(i)$ and $R_1(k)$ for some $k \in N$ and any $i \geq k$.

4.8. Lemma *Let* L *be a* C*-finite language,* X *its principal set of contexts and* $\bar{}$ *its principal mapping. Let* $w = (u, v) \in C$ *and* $x, y \in X$ *be the contexts such that* $(iL)_x \to u(iL)_y v$ *is an FG-rule for any* $i \geq d_1(L, C)$. *Then there exists* $k \geq d_1(L, C)$ *such that the following statements are equivalent*:

(i) $x \circ wRy$.

(ii) *FG-rule* $(iL)_x \to u(iL)_y v$ *is suitable for any* $i \geq k$.

Proof. (a) We prove that (i) implies (ii) for any $i \geqq d_1(L, C)$. Let $x \circ wRy$, $i \geqq d_1(L, C)$ and $t \in \{u\} (iL)_y \{v\} - (iL)_x$. Obviously the word t is of the form $t = urv$, where $r \in (iL)_y$ and $r \notin (iL)_{x \circ w}$. However $x \circ wRy$ implies $r \in L_{x \circ w}$ and consequently $|r| > i - |x \circ w|$. Thus $|t| = |urv| > i - |x \circ w| + |w| = i - |x| = c(i, (iL)_x)$ (by 4.6. (ii)).

(b) To prove that (ii) implies (i) suppose that $x \circ w\bar{R}y$. $(iL)_{x \circ w} \subseteq (iL)_y$ holds for any $i \geqq d_1(L, C)$ thus by 4.2. $L_{x \circ w}$ is finite. Let $m \geqq d_1(L, C)$ be an integer such that $L_{x \circ w} \in P(i)$ for any $i \geqq m$. Furthermore there exists $j \geqq m$ and a word r with the property $r \in (jL)_y - (jL)_{x \circ w}$ since otherwise we would have a contradiction with $x \circ w\bar{R}y$. We put $k = \max\{j, |x \circ w| + |r|\}$. For any $i \geqq k$ we have $urv \in \{u\} (iL)_y \{v\} - (iL)_x$ and $|urv| \leq i - |x| = c(i, (iL)_x)$ (by 4.6. (iii)), i.e. the FG-rule is not suitable. □

By 4.8. there exists an integer k such that for any $i \geqq k$ the sets of rules $R_1(i)$ and $R_1(k)$ coincide. The least one of these integers denoted by $d_2(L, C)$ will be caled the *second degree* of the language L.

It remains to estabilish the necessary and sufficient condition guaranteeing the coincidence of the sets $R_2(i)$ and $R_2(k)$ for some fixed $k \in N$ and any $i \geqq k$. Let L be an arbitrary language, C a finite set of nontrivial contexts. The set C is said to be complete with respect to L if there exists a nonnegative integer m such that for any context $x \in [C]$ and any word $t \in L_x$ with the property $|t| > m$ there exists a context $(u, v) \in C$ and a word $r \in V^*$ such that $t = urv$ (c.f. [10]).

4.9. Lemma *Let L be a language, C a finite set of nontrivial contexts. Let $x \in [C]$ and $t \in L_x$ be a word such that there does not exist any context $(u, v) \in C$ and a word $r \in V^*$ with the property $t = urv$. Then there exists positive integer k such that the grammar $FG(L, C, i)$ contains the rule $\overline{(iL)_x} \to t$ for any $i \geqq k$.*

Proof. We put $k = |x| + |t|$. Clearly $t \in (iL)_x$ and $t \in \overline{(iL)_x}$ for any $i \geqq k$. However $t \notin \{urv; (\overline{(iL_x)}, uQv) \in R_1(i), r \in Q\}$, thus $R_2(i)$ contains the rule $\overline{(iL)_x} \to t$ for any $i \geqq k$. □

Finally we estabilish the main theorem.

4.10. Theorem *Let L be a language, C a finite set of nontrivial contexts. Then the following statements are equivalent:*

(i) *L is FG-grammatizable,*

(ii) *L is C-finite and C is complete with respect to L.*

Proof. By 4.5. and 4.9. (i) implies (ii). Furthermore by 4.7. and 4.8. it follows that C-finiteness of the language L guarantees coincidence of the sets $R_1(i)$ and $R_1(k)$ for some fixed k and any $i \geqq k$. It remains to prove that C-finiteness of L and completeness of the set C with the respect to L guarantee coicidence of the sets $R_2(i)$ and $R_2(k)$ for some fixed $k \in N$ and any $i \geqq k$. Let X be a principal set of contexts, $^-$ a principal mapping and $k = d_2(L, C)$ the second degree of L. Completness of the set C guarantees that there exists at most finite number of the words $t \in L_x$ $(x \in [C])$ which can't be expressed in the form $t = urv$ for some $(u, v) \in C$ and by 4.9. for any word t with this property there exists $k \in N$ such that the grammar $FG(C, L, i)$ contains the rule $\overline{(iL)_x} \to t$ for any $i \geqq k$. If the grammar $FG(L, C, j)$ contains some rule $(jL)_x \to t = urv$ where $(u, v) \in C$ and $j \geqq k$, then this grammar does not contain the rule $(jL)_x \to u(jL)_y v$. By construction of the second degree of L the rule $(iL)_x \to u(iL)_y v$ is not contained in the grammar $FG(L, C, i)$ for any $i \geqq k$. By 4.8. we have $x \circ (u, v) \bar{R} y$ and consequently by 4.2 $L_{x \circ (u, v)}$ is finite since by 4.6. (ii) $(iL)_{x \circ (u, v)} \Subset (iL)_y$ holds for any $i \geq k$. Thus there exists at most finite number of the words $t = urv \in L_x$ where $x \in X$ and $(u, v) \in C$ such that the rule $(jL)_x \to t$ is contained in the grammar $FG(L, C, j)$ for some $j \geqq k$. Moreover the nonexistence of the rule $(iL)_x \to u(iL)_y u$ implies that the rule $(iL)_x \to t$ is contained in the grammar $FG(L, C, i)$ for any $i \geqq j$ and this completes the proof. □

The conditions "to be C-finite" and "to be complete" are mutualy independent (c.f. [10]).

4.11. Examples (a) Any finite language is FG-grammatizable since any set of contexts is complete with respect to any finite language and any finite language is C-finite for any set of contexts C.

(b) Any regular language is FG-grammatizable. It suffices to put $C = \{(a, \lambda); a \in V\}$. Clearly the set C is complete with respect to any language over the alphabet V and any regular language is C-finite ([3]). Moreover this construction leads to a regular grammar.

(c) Any even linear language is FG grammatizable (i.e. language generated by a grammar whose rules are either of the form $P \to vQu$ where $|u| = |v|$, or $P \to t$). We put $C = \{(a, b); a, b \in V\}$. The set C is complete with the respect to any language and any even linear language is C-finite ([10]). □

REFERENCES

[1] M. Drášil, *On languages linearly grammatizable by means of derivatives*. Arch. Math. Brno, 22, 1986, p. 139–144.

[2] R. C. Gonzales, M. G. Thomason, *Syntactic pattern recognition*. Addison–Wesley Publ. Comp., Reading, 1978.

[3] J. E. Hopcroft, J. D. Ullman, *Formal languages and their relation to automata*. Addison–Wesley Publ. Comp., Reading, 1969.

[4] B. Kříž, *Zobecněné gramatické kategorie* (*Generalized grammatical categories*). Thesis, University J. E. Purkyně, Brno, 1980.

[5] B. Kříž, *Generalized grammatical categories in the sense of Kunze*. Arch. Math., Brno, 17, 1981, p. 151 – 158.
[6] M. Novotný, *On an effective construction of a grammar generating a given language*. Prague Studies in Math., Linguistic, Prague 1983, p. 123 – 131.
[7] M. Novotný, *On some constructions of grammars for linear languages*. Intern. J. Comput. Math., 17, 1985, p. 65 – 77.
[8] M. Novotný, *Remarks on linearly grammatizable languages*. To appear in PSML 9, Prague.
[9] M. Novotný, *Personal communications*. January – May 1986.
[10] M. Novotný, *On a construction of linear grammars*. To appear in PSML 10, Prague.
[11] J. Ostravský, *Effective constructions of grammars for two particular classes*. Fundamenta informaticae 8, 1985, p. 235 – 252.
[12] K. Tanatsugu, *A grammatical inference for harmonic linear languages*. Intern. J. of Comp. and Inform. Sci., vol. 13, 5, 1984.

Milan Drášil
Czechoslovak Academy of Sciences
Institute of Geography
Mendlovo nám. 1
662 82 Brno
Czechoslovakia