On Certain Properties of Intersection Grammars

Josef Zapletal

The European Polytechnical Institut, Private University, Osvobození 699,
686 04 Kunovice, Czech Republic,
e-mail: zapletal@vos.cz

Dedicated to the Professor Noam Chomsky who combated for persecuted czech mathematics against bolshevik injury in ČSSR for a long time on radio stations Free Europe, BBC London and Voice of America.

Abstract: This paper links up to the papers [1], [2] and [3]. The author deals with the intersections of languages of various types (in the sense of Chomsky) and studies two types of resultand languages. It is proved that the intersection of two languages of the type 2 can be a language of the type 1 and otherwise the intersection of two languages of the type 1 is a language of the type 3. The known definitions and theorems are given in the introductionary chapter without proofs.

Key words. V^* the free monoid over the set V, algebraic language, Chomsky Grammars, x-derivative of the element y, the rules of grammar, terminal and non-terminal elements, the clause and structural grammars, the class of languauges, the intersection grammars.

1. Preliminaries

1.1 Definition Let V be a set. Then arbitrary element $r = (x, y) \in V^* \times V^*$ is called a *rule* over the set V. The string y is caled $left$ and the string x *right side* of the rule r.

1.2 Definition Let V be a set, $s \in V^*, t \in V^*$ strings, $(x, y) \in V^* \times V^*$ a rule. We say that the string t was $infered$ from the string s by *the utilizing of the rule* (y, x) and we write $s \Rightarrow t(\{y, x\})$ if there exist strings $u \in V^*, v \in V^*$ so that $s = uyv, uxv = t$.

1.3 Definition Let V be a set, $R \subseteq V^* \times V^*$ the set of rules. $s \in V^*, t \in V^*$ strings. We put $s \Rightarrow t\,(R)$ if there exists $(x, y) \in R$ such that $s \Rightarrow t(\{y, x\})$ Then we say that the string t is $direct\ infered$ from the string s *utilizing of the rules of the set* R.

1.4 Definition Let V be a set,$R \subseteq V^* \times V^*$ the set of rules. $s \in V^*, t \in V^*$ strings. Let m be a positive integer and $\{s_i\}_{s=0}^{m}$ be a finite sequence of strings from the set V^* such that $s = s_0, s_{i-1} \Rightarrow s_i\,(R)$ for every i with the property $1 \leq i \leq m$ and $s_m = t$. Then we say that the sequence $\{s_i\}_{s=0}^{m}$ is $the\ s - derivative\ of\ the\ string\ t\ in\ the\ set\ R\ of\ the\ length\ m$.

1.5 Remark For $m = 0$ the sequence $\{s_i\}_{s=0}^{m}$ from 1.4 has exactly one member $s = s_0 = t$. We say that this one-member sequence $\{s_i\}_{s=0}^{0}$ is *a trivial* s-derivative of the string t.

1.6 Definition Let V be a set, $R \subseteq V^* \times V^*$ the set of rules. $s \in V^*, t \in V^*$ strings. We say that the string t can be infered from the string s and we write $s \overset{*}{\Rightarrow} t\,(R)$ if there exists at least one s derivative of the string t in the set R.

1.7 Definition We suppose that a set U and its finite subset V is given. We call the elements from the set V by *terminal symbols*, the elements from the set $U - V$ by $non - terminal\ symbols$. Further let $R \subseteq U^* \times U^*$ be arbitrary set of rules and a special set $S \subseteq U^*$ which elements are called *initial strings*. The ordered quaternion $G =< U, V, S, R >$ is called *generalized grammar*. We put

$$L(G) = \{w \in V^*; \text{ there exists } s \in S \text{ such that } s \overset{*}{\Rightarrow} w\,(R)\}$$

The pair $(V, L(G))$ is called the *language generated by generalized grammar* G

1.8 Definition Generalized grammar $G =< U, V, S, R >$ is called *the grammar* when the sets U, S and R are finite.

1.9 Theorem [4] Let V be nonvoid finite set, W countable set, let $V \cap W = \emptyset$ Then the system of all the grammars of the form $< U, V, S, R >$ where $U = V \cup Z$ for some finit $Z \subseteq W$ where further $S \subseteq U^*$ and $R \subseteq U^* \times U^*$ is countable.

1.10 Corollary [4] Let V be a nonvoid finite set. Then the set of all languages of the form (V, L) which can be generated by grammars is countable.

1.11 Theorem [4] Let V be a nonvoid finite set. Then the set of all languages of the form (V, L) is uncountable.

1.12 Theorem [4] Let V be a nonvoid finite set. Then the set of all languages of the form (V, L) which can be not generated by any grammar is uncountable.

2. The Chomsky hierarchy of grammars and languages.

2.1 Definition Let $G =< U, V, S, R >$ be a grammar. It is called *the the type* $-$ 0 *grammar* if the following conditions are contend:

(A) There exists $s \in U - V$ such that $S = \{s\}$.
(B) For every $(y, x) \in R\ \ y \in U^*(U - V)U^*$ holds true.

2.2 Remark It is proved that arbitrary language is generated by grammer exactly when it is generated by grammar of the type 0.

2.3 Definition Let $G =< U, V, S, R >$ be a type-0 grammar . We say that it is of *the type* 1 if the following conditions are satisfied:

(C) The initial symbol is not contained in x for any $(y, x) \in R$.
(D) The inequality $\mid y \mid \leq \mid x \mid$ holds true for every $(y, x) \in R$ where y is different from the initial symbol.

2.4 Remark In the grammar of the type 1 the initial symbol can occur only in the left sides of the rules. The right sides are non void with only one exception which creates the rule (s, λ) in which s is the initial symbol.

2.5 Definition Let $G =< U, V, S, R >$ be a grammar of the type 1. We say that it is *of the type* 2 if the following condition is satisfied:

(E) It holds $y \in U - V$ for every $(y, x) \in R$.

2.6 Remark The left side of arbitrary rule of the grammar of the type 2 is formed by one nonterminal symbol. Grammars of this type are called noncontext grammars often.

2.7 Definition Let $G =< U, V, S, R >$ be a grammar of the type 2 where s is initial symbol and $S = \{s\}$. We say that that this grammar is *of the type* 3 if the following condition is satisfied:

(F) If $(y, x) \in R$ and (s, λ) then either $x \in V$ nor $x = vu$ where $v \in V$ and $u \in U - V$

2.8 Theorem If $i \in \{1, 2, 3\}$ then every grammar of the type i is simultaneously the grammar of the type $i - 1$.

2.9 Definition Let be $i \in \{0, 1, 2, 3\}$. The language is called *the language of the type* i if there exist a grammar of the type i which them generates

2.10 Theorem Let be $i \in \{0, 1, 2, 3\}$. Then every language of the type i is simultaneously the language of the type $i - 1$.

2.11 Remark It may be that one and the same languge can be generated by different grammars. These grammars can be of different types.

3. The constructions of intersection grammars

3.1 Example The intersection of two languages of the type 2 is not a language of the type 2 in full generality.

Proof. Let (V_1, L_1) be the language generated by the grammar $G_1 =< U_1, V_1, S_1, R_1 >$ where $U_1 = \{S_1, P, T, a, b\}$ where S_1 is the initial symbol, and $V_1 = \{a, b\}$ the rules are:

$$\begin{aligned} R_1 : S_1 &\to PT, \\ P &\to aPb, \\ T &\to TT, \\ P &\to ab, \\ T &\to a. \end{aligned}$$

Analogously let (V_2, L_2) be the language generated by the grammar $G_2 =< U_2, V_2, S_2, R_2 >$ where $U_2 = \{S_2, M, N, a, b\}$ where S_2 is the initial symbol, and $V_2 = \{a, b\}$ the rules are:

$$R_2 : S_2 \to NM,$$

$$M \to bMa,$$
$$N \to NN,$$
$$M \to ba,$$
$$N \to a$$

The both grammars are of the type 2.The language V_1, L_1 generated by the grammar G_1 contains the sentences of th form $a^n b^n a^r$, where $n, r = 1, 2, \ldots$. and the language generated by the grammar G_2 contains the sentences of the form $a^r b^n a^n$ where $n, r = 1, 2, \ldots$. We create the language $(V_1, L_1(G_1)) \cap (V_2, L_2(G_2))$.

At first we create the new grammars $\bar{G}_1, \bar{G}_2$ as the modification of G_1 and G_2. We rewrite the alphabet of G_1 as follows: we put $a = \alpha_1, b = \alpha_2, P = \alpha_3, T = \alpha_4, S_1 = \alpha_5$ such that $\bar{V}_1 = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$ is full vocabulary of $\bar{G}_1$. Analogous to G_1 we rewrite the alphabet of G_2. We receive $a = \beta_1, b = \beta_2, M = \beta_3, N = \beta_4, S_2 = \beta_5$ such that $\bar{V}_2 = \{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ is full vocabulary of $\bar{G}_2$. We enter further set of symbols $\Gamma = \{\gamma_{i,j}\}$ for $i, j = \{1, 2, 3, 4, 5\}$ and we put $\gamma_{ij} = \{\alpha_i, \beta_j\}$. Now we define the intersection grammar G:

$G = (\bar{V}_1 \cup \Gamma \cup \{a, b\} \cup \gamma_{55}, R)$ where γ_{55} defines the initial symbol and system of rules R is devided into four families.

R :

I
1) $\alpha_5 \to \alpha_3\alpha_4$
2) $\alpha_3 \to \alpha_1\alpha_3\alpha_2$
3) $\alpha_4 \to \alpha_4\alpha_4$
4) $\alpha_3 \to \alpha_1\alpha_2$
5) $\alpha_4 \to \alpha_1$
6) $\gamma_{55} \to \alpha_3\gamma_{45}$
7) $\gamma_{35} \to \alpha_1\alpha_3\gamma_{25}$
8) $\gamma_{45} \to \alpha_4\gamma_{45}$
9) $\gamma_{35} \to \alpha_1\gamma_{25}$
10) $\gamma_{45} \to \gamma_{15}$

II
1) $\alpha_{k_1}\gamma_{k_2 5} \to \gamma_{k_1 4}\gamma_{k_1 3}$
2) $\alpha_{k_1}\alpha_{k_2}\gamma_{k_3 3} \to \gamma_{k_1 2}\gamma_{k_1 3}\gamma_{k_3 1}$
3) $\alpha_{k_1}\gamma_{k_2 4} \to \gamma_{k_1 4}\gamma_{k_2 4}$
4) $\alpha_{k_1}\gamma_{k_2 3} \to \gamma_{k_1 2}\gamma_{k_2 1}$
5) $\gamma_{k_1 4} \to \gamma_{k_1 1}$
where $k_1, k_2, k_3 = 1, 2$

III
$\gamma_{ij}\alpha_k \to \alpha_i\gamma_{kj}$ $i, k = 1, 2$
$\alpha_i\gamma_{kj} \to \gamma_{ij}\alpha_k$ $j = 1, 2, 3, 4, 5$

IV
$\gamma_{11} \to a$
$\gamma_{22} \to b$

Such constructed grammar is the grammar of the language $(V_1, L_1(G_1)) \cap (V_2, L_2(G_2))$ wchich is of the type 1 and contains the sentences $a^m b^m a^m$ and only such sentences. We utilize the intersection grammar G for derivation of the sentences of the language L.

The utilizing of the fourth rule from the first family on the string xyz with the result uwv wil be described as followes: $xyz \overset{I/4}{\Rightarrow} uwv$, optionally multiple application of rules from the part III as $xyz \overset{III*}{\Rightarrow} uwv$.

1) $\gamma_{55} \overset{I/6}{\to} \alpha_3\gamma_{45} \overset{I/4}{\Rightarrow} \alpha_1\alpha_2\gamma_{45} \overset{I/10}{\Rightarrow} \alpha_1\alpha_2\gamma_{15} \overset{II/1}{\Rightarrow} \alpha_1\gamma_{24}\gamma_{13} \overset{III}{\Rightarrow} \gamma_{14}\alpha_2\gamma_{13} \overset{II/5}{\Rightarrow} \gamma_{11}\alpha_2\gamma_{13} \overset{II/4}{\Rightarrow} \overset{II/4}{\Rightarrow} \gamma_{11}\gamma_{22}\gamma_{11} \overset{IV*}{\Longrightarrow} aba.$

2) $\gamma_{55} \overset{I/6}{\to} \alpha_3\gamma_{45} \overset{I/2}{\Rightarrow} \alpha_1\alpha_3\alpha_2\gamma_{45} \overset{I/4}{\Rightarrow} \alpha_1\alpha_1\alpha_2\alpha_2\gamma_{45} \overset{I/8}{\Rightarrow} \alpha_1\alpha_1\alpha_2\alpha_2\alpha_4\gamma_{45} \overset{I/5}{\Rightarrow} \alpha_1\alpha_1\alpha_2\alpha_2\alpha_1\gamma_{45} \overset{I/10}{\Rightarrow}$

$\overset{I/10}{\Rightarrow} \alpha_1\alpha_1\alpha_2\alpha_2\alpha_1\gamma_{15} \overset{II/1}{\Rightarrow} \alpha_1\alpha_1\alpha_2\alpha_2\gamma_{14}\gamma_{13} \overset{III*}{\Longrightarrow} \alpha_1\gamma_{14}\alpha_2\alpha_2\alpha_1\gamma_{13} \overset{II/2}{\Rightarrow} \alpha_1\gamma_{14}\alpha_2\gamma_{22}\gamma_{13}\gamma_{11} \overset{III}{\Rightarrow}$

$\overset{III}{\Rightarrow} \alpha_1\gamma_{14}\gamma_{22}\alpha_2\gamma_{13}\gamma_{11} \overset{II/4}{\Rightarrow} \alpha_1\gamma_{14}\gamma_{22}\gamma_{22}\gamma_{11}\gamma_{11} \overset{II/3}{\Rightarrow} \alpha_1\gamma_{14}\gamma_{22}\gamma_{22}\gamma_{11}\gamma_{11} \overset{II/3}{\Rightarrow} \gamma_{14}\gamma_{14}\gamma_{22}\gamma_{22}\gamma_{11}\gamma_{11} \overset{II/5*}{\Longrightarrow}$

$\overset{II/5*}{\Longrightarrow} \gamma_{11}\gamma_{11}\gamma_{22}\gamma_{22}\gamma_{11}\gamma_{11} \overset{IV*}{\Longrightarrow} aabbaa = a^2b^2a^2$

We constract a new grammar $\mathcal{G}$ specialized in the language with the sentences $a^nb^na^n$.with the alphabet of nonterminals $\mathcal{U} - \mathcal{V} = \{S, A, B, C, D\}$ where S is the initial symbol, with the alphabet of terminal symbols $\mathcal{V} = \{a, b\}$, database vocabulary $\mathcal{U}$ and with the rules

R: 1) $S \to ABC$, 2) $AB \to AAD$,
3) $DC \to BBCC$, 4) $DB \to BD$,
5) $A \to a$ 6) $aB \to ab$
7) $bB \to bb$ 8) $bC \to bc$
9) $cC \to cc$

what can be written $\mathcal{G} = (\mathcal{U}, \mathcal{V}, S, R)$. We show that the derivations building by the grammar $\mathcal{G}$ is more simple than the building by the intersection grammar G. We infer also the sentences aba and $a^2b^2a^2$.

$S \overset{1}{\to} ABC \overset{5,6,7,8,9*}{\Longrightarrow} abc$
$S \overset{1}{\to} ABC \overset{2}{\Rightarrow} AADC \overset{3}{\Rightarrow} AABBCC \overset{5,6,7,8,9*}{\Longrightarrow} a^2b^2a^2$

A more complex is the derivation of the sentence $a^3b^3a^3$ We aply also the fourth rule in this derivation. We receive:

$S \overset{1}{\to} ABC \overset{2}{\Rightarrow} AADC \overset{3}{\Rightarrow} AABBCC \overset{2}{\Rightarrow} AAADBCC \overset{4}{\Rightarrow} AAABDCC \overset{3}{\Rightarrow}$
$\overset{3}{\Rightarrow} AAABBBCCC \overset{5,6,7,8,9*}{\Longrightarrow} a^3b^3c^3$

3.2 Assertion The language $a^nb^na^n$ is not the language of the type 2.

Proof. We suppose that G is a type 2 grammar of the language $(V, L) = a^nb^na^n$.Language (V, L) contais infinitaly many sentences, but the grammar G has a finite vocabulary. Therefore we can find an $A \in U - V$ which is included in infinitely many terminal derivatives in their penultimate string. There exists a string of maximal length z conteining only terminal symbols among the the set of rules of the form $A \to v$ which is finite. Let be $\mid z \mid = q$ where q is natural number. There exist out of number terminal stringsfor which $A \overset{*}{\Rightarrow} w$ and $\mid w \mid > q$
Let us suppose the sentence $a^nb^na^n$ such that $n > q$ .This sentence arose from the string xAy using the rule $A \to z$. Appliing on the string xAy overwrite mode $A \overset{*}{\Rightarrow} w$ insted of $A \to z$ we receive sentences which do not attache to the language (V, L). In particular: The penultimate string ahead of the application of the rule $A \to z, \mid z \mid < q$ acquires some of the following forms:

1) $xAy = a^sAa^tb^na^n$, 4) $xAy = a^nb^sAa^t$,
2) $xAy = a^sAb^ta^n$, 5) $xAy = a^nb^na^sAa^t$,
3) $xAy = a^nb^sAb^ta^n$, where $s, t = 0, 1, \ldots n$ and $A^0 = B^0 = \lambda$.

We choose some of overwrites $A \overset{*}{\Rightarrow} w$ where $\mid w \mid > q$ and we receive strings which do not

belong to the language (V, L) generated by intersection grammar G. Hence the grammar of the type 2 does not generate the language with sentences of the form $a^n b^n a^n$.

3.3 Remark There exist languages $(V, L_1), (V, L_2)$ of the type 1 for whose the intersection language $(V, L_1) \cap (V, L_2)$ is of the type 3.
Let $(V, L_1) = \{a^m b^n a^m b^n ccc\}$, $m, n \geq 1$ and $(V, L_2) = \{ccca^m b^m a^m b^n \cup ababccc\}$, $m, n \geq 1$ both this laqnguages are of the type 1 [1]. The intersection language $(V, L_1) \cap (V, L_2)$ contains one and only one sentence *ababccc* which can be generated by grammar $G =< U, V, S, R >$, where $U = \{A, B, C, D, M, N, S, a, b, c\}$, $V = \{a, b, c\}$, S is the initial symbol and

$$
\begin{array}{lll}
R: & S \to aA & D \to cM, \\
 & A \to bB & M \to cN, \\
 & B \to aC & N \to c, \\
 & C \to bD &
\end{array}
$$

The system of rules R of the grammar G satisfies the condition F from the definition 2.7. Hence G is of the type 3 and the intersection language (V, L) is also of the type 3.

References

[1] CHOMSKY N.: On Certain Formal Properties of Grammars. Information and Control **2** , 137-167 (1959)

[2] CHOMSKY N., Miller G.A.: Introduction to the Formal Analysis of Natural Languages. Handbook of Mathematical Psychology, Vol.**2**, New York, Wiley, 1963, 269-322

[3] ČULÍK K. : Some Notes of Finite State Languages and Events Represented by Finite Automata Using Labelled Graphs. Časopis pro pěstování matematiky 86 (1961) Praha 43-55.

[4] ČULÍK K. : Axiomatic System for Phrase Structure Grammars. I Information and Control. **8** (1965) 493-502.

[5] ČULÍK K. : On Languages Generated by Some Types of Algorithms. Prague Studies in Mathematical LinguisticsPublishing House of the Czechoslovak Academy of Sciences Prague 1966, 141-146.

[6] LANDWEBER P.S.: Three Theorems on Phrase Structure Grammars of Type 1. Information and control **6**,1963 131-136.

[7] NOVOTNÝ M.: Über endlich Charakterisierbare Sprachen. Publ. Fac. Sci. Univ. J.E Purkyně Brno 1965 N^o 468, 495-502.

[8] NOVOTNÝ M.: On Some Operators reducing Generalized Grammars.Information and Control 26 (1974), 225-235.

[9] NOVOTNÝ M.: Construction of Pure Grammars. Fundamenta Informaticae 52 (2002), 345-360

[10] SCHÜTZENBERGER M.P. On an Application of Semigroups Methods to Some Problems in Coding. IRE Transactions on Information Theory. V. IT-2, N^o 3,1956 47-60.

Doc.RNDr.Josef Zapletal, CSc.
Evropský polytechnický institut,
soukromá vysoká škola,
Osvobození 399, 686 04 Kunovice
E-mail epi@vos.cz
Telefon: +420 573 548 035