NONTERMINALS AND CODINGS IN DEFINING VARIATIONS OF OL-SYSTEMS

Sven Skyum
Department of Computer Science
University of Aarhus
Aarhus, Denmark

Summary

The use of nonterminals versus the use of codings in variations of OL-systems is studied. It is shown that the use of nonterminals produces a comparatively low generative capacity in deterministic systems while it produces a comparatively high generative capacity in nondeterministic systems.

Finally it is proved that the family of context-free languages is contained in the family generated by codings on propagating OL-systems with a finite set of axioms, which was one of the open problems in [10]. All the results in this paper can be found in [71] and [72].

1. Definitions

By definition, an EOL-system is a quadruple $G = \langle \Sigma, P, \omega, \Delta \rangle$, where Σ and Δ are alphabets with $\Delta \subseteq \Sigma$, P is a finite set of context-free productions containing at least one production for every letter of Σ, and $\omega \in \Sigma^+$. The direct yield relation $\Rightarrow$ on the set Σ^* is defined as follows: $x \Rightarrow y$ holds iff there is an integer $k \geqq 1$, letters a_i and words α_1, $1 \leqq i \leqq n$, such that

$$x = a_1 \ldots a_n, \; y = \alpha_1 \ldots \alpha_n,$$

and $a_i \rightarrow \alpha_i$ is a production in P, for each $i = 1, \ldots, n$. The relation $\overset{*}{\Rightarrow}$ is the reflexive transitive closure of $\Rightarrow$. The language L(G) generated by G is defined by

$$L(G) = \{ w \in \Delta^* \mid \omega \overset{*}{\Rightarrow} w \}.$$

The EOL-system is an OL-system iff $\Delta = \Sigma$. It is deterministic (abbreviated: D) iff there is exactly one production for every letter of Σ. It is propagating (abbreviated: P) iff the right side of every productions is distinct from the empty word λ. We may also combine these notions and speak, for instance, of PDOL- or EPOL-systems.

We also consider generalizations of the systems defined above obtained by replacing the axiom ω by a finite set Ω of axioms. The language generated by such a system consists of the (finite) union of the languages generated by the systems obtained by choosing each element $\omega \in \Omega$ to be the axiom. This generalization is denoted by the letter F. Thus, we may speak of EPDFOL-systems.

For any class of systems, we use the same notation for the family of languages generated by these systems. E.g., EPDOL denotes the family of languages generated by EPDOL-systems.

By a coding we mean a length-preserving homomorphism (often also called a literal homomorphism). The prefix C attached to the name of a language family indicates that we are considering codings of the languages in the family.

2. Deterministic systems

We start by examining the relation between the use of nonterminals and codings in deterministic systems.

Theorem 2.1

EDOL $\subsetneq$ CDOL and EPDOL $\subsetneq$ CPDOL.

Proof

We will only prove the first inclusion. The second one is proved in the same way.

Now let $G = \langle \Sigma, P, \omega, \Delta \rangle$ be an EDOL-system. The following describes the construction of a DOL-system $G = \langle \Sigma', P', \omega', \Sigma' \rangle$ and a coding h from Σ' into Δ such that $L(G) = h(L(G'))$. For a word x, $\min(x)$ denotes the set of letters occurring in x..

Consider the sequence of words from Σ^* generated by $G, \omega = \omega_1, \omega_2, \omega_3, \ldots$. There exist natural numbers n and m such that $\min(\omega_m) = \min(\omega_{m+n})$, which implies that for any $i \geq 0$ and any j, $0 \leq j < n$:

$$\min(\omega_{m+j}) = \min(\omega_{m+ni+j}). \tag{1}$$

Let d_k denote the cardinality of $\min(\omega_k)$, $1 \leq k < m+n$. Define

$$N_\Delta = \{k \in N \mid 1 \leq k < m+n,\ \min(\omega_k) \subseteq \Delta\}.$$

For any $k \in N_\Delta$ introduce new symbols not in Σ

$$\Sigma_k = \{ a_{kj} \mid 1 \leq j \leq d_k \},$$

and define isomorphism f_k mapping $\min(\omega_k)$ onto Σ_k, where

$$f_k(a) = a_{kj} \text{ iff } a \text{ is the j'th symbol of } \min(\omega_k),\ k \in N_\Delta$$

Note that the f_k's are defined for some fixed enumerations of the sets $\min(\omega_k)$. Σ' is going to be the union of the above defined Σ_k's:

$$\Sigma' = \bigcup_{k \in N_\Delta} \Sigma_k \quad .$$

Define k_1 and k_2 as the minimal and maximal elements of N_Δ.

For any of the letters a_{kj} where $k \neq k_2$ define production in P':

$$a_{kj} \to f_{k'}(\alpha)$$

where k' is the smallest element in N_Δ greater than k and α is the string derived from $f_k^{-1}(a_{kj})$ in (k'-k) steps in G.

It follows from (1) that L(G) is finite if $k_2 < m$. If this is the case then define for any j, $1 \leq j \leq d_{k_2}$ production in P':

$$a_{k_2 j} \to a_{k_2 j}.$$

Otherwise, let k_3 denote the minimal element in N_Δ greater than or equal to m, and define productions in P' for any j, $1 \leq j \leq d_{k_2}$, for which $f_{k_2}^{-1}(a_{k_2 j})$ derives some string $\alpha \in \Delta^*$ in $(n-k_2+k_3)$ steps in G:

$$a_{k_2} j \to f_{k_3}(\alpha) \; .$$

Note that the use of f_{k_3} is well defined since $\min(\omega_{k_2+(n-k_2+k_3)}) = \min(\omega_{k_3})$ (from the fact that $k_3 \geq m$ and (1) above). Finally define the coding h from Σ' into Δ in the way that for every $a_{kj} \in \Sigma'$: $h(a_{kj}) = f_k^{-1}(a_{kj})$. Then

$$L(G) = h(L(G'))$$

where $G' = \langle \Sigma', P', f_{k_1}, (\omega_{k_1}), \Sigma' \rangle$, and this proves the inclusion of the theorem. The inclusion is proper because $\{a^n \mid n \geq 1\} \in$ CDOL\EDOL.

We have the following theorem as an immediate consequence of theorem 2.1.

Theorem 2.2

EDFOL $\subsetneq$ CDFOL and EPDFOL $\subsetneq$ CPDFOL.

3. Nondeterministic systems

We will now examine the relations between the nondeterministic families, corresponding to those occurring in Theorems 2.1 and 2.2.

The following two theorems correspond to Theorem 2.1. The proof of the first one can be found in [20].

Theorem 3.1

COL = EOL.

Theorem 3.2

CPOL $\subsetneq$ EPOL.

Proof

The inclusion is easily checked. The inclusion is proper because the language $L = \{a^n b^n c^n \mid n \geq 1\} \cup \{d^{3^n} \mid n \geq 1\}$ belongs to EPOL, but is does not belong to CPOL. The proof for the later statement can be found in [10].

Notice here that while the generating capacity due to the use of nonterminals was weaker than the generating capacity due to codings in deterministic propagating OL-systems, the converse is true if you are dealing with nondeterministic propagating OL-systems.

The following theorem corresponds to Theorem 2.2. The proof can be found in [72].

Theorem 3.3

CFOL = EFOL and CPFOL $\subseteq$ EPFOL.

It is an open problem whether or not the inclusion CPFOL $\subseteq$ EPFOL is proper.

A somewhat related problem is whether or not the family of context-free languages is included in the family CPFOL. Indeed we have the following theorem.

Theorem 3.4

The family of context-free languages is properly included in the family CPFOL.

Proof

Let $G = \langle V, \Sigma, P, S\rangle$ be a cf-grammar of a language not containing λ in Greibach-normal form (i.e., the productions are of the form $A \rightarrow a$ or $A \rightarrow aA_1 \ldots A_n$). Suppose there are no useless symbols in V.

For each $A \in V$ we choose

$$w_A \in \{w \in \Sigma^* \mid A \overset{*}{\Rightarrow} w, \ |w| \text{ minimal}\}.$$

w_A will, in the rest of the proof, be fixed for every letter $A \in V$.

Define $k : V \rightarrow N$ by $k(A) = |w_A|$, and furthermore

$$s(A) = \{ xw_x \in \Sigma^{k(A)}V^* \mid A \overset{*}{\underset{left}{\Rightarrow}} xw_x,\ |x| = k(A) \} \text{ and}$$

$$m(A) = \{ x \in \Sigma^{k(A)} \mid \exists\ w \in V^* : xw \in s(A) \}$$

Since the grammar was in Greibach normal form, $s(A)$ and $m(A)$ are finite sets of strings.

Let $n : V \to N$ be defined as $n(A) = \{$number of strings in $m(A)\}$.

We will use $m(A)$ as an ordered set.

Now we can construct a PFOL-system H and a coding h such that $h(L(G)) = L(G)$:

$$H : < \Sigma \cup \bigcup_{\substack{A \in V \\ 1 \le i \le k(A) \\ 1 \le j \le n(A)}} A_i^j,\ P',\ \{ S_1^1 S_2^1 \ldots S_{k(S)}^1, S_1^2 S_2^2 \ldots S_{k(S)}^2, \ldots, S_1^{n(S)} S_2^{n(S)} \ldots S_{k(S)}^{n(S)} \} >$$

P' is defined **as** follows:

1) For all $a \in \Sigma$, $a \to a$ is in P'.
2) For all $A \in V$, $1 \le j \le n(A)$, and $1 \le i \le k(A)-1$, $A_i^j \to a_i^j$ is in P', where a_i^j is the i'th terminal in the j'th string in $m(A)$.
3) For all $A \in V$ and $1 \le j \le n(A)$

$$A_{k(A)}^j \to a_{k(A)}^j B_{11}^{k_1} B_{12}^{k_1} \ldots B_{1k(B_1)}^{k_1} B_{21}^{k_2} B_{22}^{k_2} \ldots B_{2k(B_2)}^{k_2} \ldots$$

$$\ldots B_{q1}^{k_q} B_{q2}^{k_q} \ldots B_{qk(B_q)}^{k_q}$$

is in P' for all $B_1, B_2, \ldots, B_q$ and $1 \le k_i \le n(B_i)$ where $xB_1B_2 \ldots B_q \in S(A)$ and x is the j'th string in $m(A)$.

The coding h is defined by $h(a) = a$ for all $a \in \Sigma$, and $h(A_1^j A_2^j \ldots A_{k(A)}^j) = w_A$, for all $A \in V$ and $1 \le j \le n(A)$.

We prove that $L(G) \subseteq h(L(H))$. The other inclusion is shown in the same way.

Let $w \in L(G)$.
There exists a derivation of w in G such that

$$S = A_1 \overset{*}{\underset{left}{\Rightarrow}} x_1' A_2 A_3 \ldots A_n$$

$$\overset{*}{\Rightarrow} x_1' x_2' B_{21} \ldots B_{2q_2} x_3' B_{31} \ldots B_{3q_3} \ldots x_n' B_{n1} \ldots B_{nq_n}$$

$$\overset{*}{\Rightarrow} x_1' x_2' x_{21}'' \ldots x_{2q_2} x_3' x_{31}'' \ldots x_{3q_3}'' \ldots x_n' x_{n1}'' \ldots x_{nq_n}'' = w$$

where $x_i' \in m(A_i)$ for all $1 \leq i \leq n(A)$ and $B_{ij} \overset{*}{\Rightarrow} x_{ij}''$ for $2 \leq i \leq n$ and $1 \leq j \leq q_i$.

It suffices then to show that there exists an axiom $S_1' S_2' \ldots S_{k(S)}'$ in H such that:

$$S_1' S_2' \ldots S_{k(S)}' \underset{H}{\Rightarrow}$$
$$x_1' A_{21}^{k_2} A_{22}^{k_2} \ldots A_{2k(A_2)}^{k_2} A_{31}^{k_3} A_{32}^{k_3} \ldots A_{3k(A_3)}^{k_3} \ldots A_{n1}^{k_n} A_{n2}^{k_n} \ldots A_{nk(A_n)}^{k_n}$$

and

$A_{j1}^{k_j} A_{j2}^{k_j} \ldots A_{jk(A_j)}^{k_j} \underset{H}{\Rightarrow} x_j' w_j$ for all $2 \leq j \leq n$ but that is exactly how H is constructed.