START STRING IN FORMAL LANGUAGE THEORY

Lukáš Rychnovský
Doctoral Degree Programme (2), FIT BUT
E-mail: rychnov@fit.vutbr.cz

Supervised by: Alexander Meduna
E-mail: meduna@fit.vutbr.cz

ABSTRACT

This article introduces the notion of start string to formal language theory. It is usual to start derivation from single nonterminal but in this article we study grammars where the derivation start from string of nonterminals. This approach leads to infinite language hierarchy according to length of start string.

1 INTRODUCTION

In classic formal language theory it is usual to start derivation from single nonterminal. It is obvious because all grammar classes in Chomsky hierarchy generate the same language even if they start from a string of nonterminals. In this article we present a right-linear grammar with start string regulated by regular language which generate different languages when it starts from strings of different lengths. Moreover these languages form infinite hierarchy according to the length of their start string.

2 START STRING

Definition 2.1. *Let $n \geq 1$. A* right-linear grammar with a start string of length n*, n-RLG for short, is a quadruple $G = (N,T,R,S)$, where N and T are alphabets such that $N \cap T = \emptyset, S \in N^+, |S| \leq n$, and R is a finite set of productions of the form $A \to x$, where $A \in N$ and $x \in T^*(N \cup \{\varepsilon\})$. Set $V = T \cup N$.*

Let Ψ be an alphabet of rule labels such that card(Ψ) = card(R), and ψ be a bijection from R to Ψ. For simplicity, to express that ψ maps a rule $A \to x \in R$, to ρ, where $\rho \in \Psi$, we write $\rho.A \to x \in R$; in other words, $\rho.A \to x$ means $\psi(A \to x) = \rho$.

If $\rho.A \to x \in R$ and $u, v \in V^*$, then we write $uAv \Rightarrow uxv\ [\rho]$ in G.

Let $\chi \in V^*$. Then G makes the zero-step derivation from χ to χ according to ε, symbolically written as $\chi \Rightarrow^0 \chi\ [\varepsilon]$. Let there exist a sequence of derivation steps $\chi_0, \chi_1, \ldots, \chi_n$ for some $n \geq 1$ such that $\chi_{i-1} \Rightarrow \chi_i\ [\rho_i]$, where $\rho_i \in \Psi$, for all $i = 1, \ldots, n$, then G makes n derivation steps from χ_0 to χ_n according to $\rho_1 \ldots \rho_n$, symbolically written as $\chi_0 \Rightarrow^n \chi_n\ [\rho_1 \ldots \rho_n]$. If for some $n \geq 0, \chi_0 \Rightarrow^n \chi_n\ [\rho]$, where $\rho \in \Psi^*$ and $|\rho| = n$, we write $\chi_0 \Rightarrow^* \chi_n\ [\rho]$.

We call a derivation $S \Rightarrow^* w$ *successful*, if and only if, $w \in T^*$.

Let Ξ be a control language over Ψ; that is, $\Xi \in \Psi^*$.

Under the regulation by Ξ, the language that G generates is denoted by $L(G,\Xi)$ and defined as

$$L(G,\Xi) = \{w \mid S \Rightarrow^* w\ [\rho], \rho \in \Xi, w \in T^*\}.$$

Let i be a positive integer and X be a family of languages. Set

$$\Re(X,i) = \{L \mid L = L(G,X), \text{where } G = (N,T,R,S) \text{ is a } i\text{-RLG}\}.$$

Specifically, $\Re(REG,i)$ is central to this paper, where REG denotes the family of regular languages.

Definition 2.2. *Let $G = (N,T,R,S)$ be an n-RLG for some $n \geq 1$ (See Definition 2.1). $G = (N_1,N_2,\ldots,N_n,T,R_1,R_2,\ldots,R_n,S)$ is a* distributed n-RLG*, $n-_{dis}RLG$ for short, if*

- *$N = N_1 \cup N_2 \cup \ldots \cup N_n$, where $N_i, 1 \leq i \leq n$ are pairwise disjoint nonterminal alphabets,*
- *$S = X_1X_2\ldots X_n$, $X_i \in N_i, 1 \leq i \leq n$,*
- *$R = R_1 \cup R_2 \cup \ldots \cup R_n$,*
  such that for every $A \rightarrow xB \in R_i$, $A,B \in N_i$, for some $1 \leq i \leq n, x \in T^*$
  and for every $A \rightarrow a \in R$, $A \in N, a \in T^*$.

Set $\Psi_i = \{\rho \mid \rho.A \rightarrow aB \in R_i$ or $\rho.A \rightarrow a \in R_i$, where $A,B \in N_i, a \in T^*\}$.

Theorem 2.1. *For all $n \geq 1$, $\mathcal{L}(n-_{dis}RLG) = \mathcal{L}(n\text{-}RLG)$.*

Proof. See [Rych–08]. □

Definition 2.3. *Let $i \geq 1$ and X be a family of languages. Let $L(G,\Xi)$ be a language generated by G and regulated by Ξ (See definition 2.1). Set*

- *$\Re(X,i) = \{L \mid L = L(G,\Xi)$, where $G = (N,T,R,S)$ is a i-RLG and $\Xi \in X\}$.*
- *${}_{dis}\Re(X,i) = \{L \mid L = L(G,\Xi)$, where $G = (N_1,N_2,\ldots,N_n,T,R_1,R_2,\ldots,R_n,S)$ is a $i-_{dis}RLG$ and $\Xi \in X\}$.*

Definition 2.4. *(See [Wood–73]) For $n \geq 1$, an* n-parallel right-linear grammar*, n-PRLG for short, is an (n+3)-tuple $G = (N_1,\ldots N_n,T,S,P)$ where*

- *$N_i, 1 \leq i \leq n$ are pairwise disjoint nonterminal alphabets,*
- *T is a terminal alphabet, $N \cap T = \emptyset$,*
- *$S \notin N_1 \cup \ldots \cup N_n$ is the start symbol,*
- *P is a finite set of rules. P contains three kinds of rules*

  1. *$S \rightarrow X_1\ldots X_n, \quad X_i \in N_i, 1 \leq i \leq n$,*

2. $X \to aY, \quad X,Y \in N_i$, *for some* $1 \le i \le n, a \in T^*$, *and*

3. $X \to a, \quad X \in N_i$, *for some* $1 \le i \le n, a \in T^*$.

For $x,y \in (N \cup T \cup \{S\})^*, x \Rightarrow y$ *if and only if*

- *either* $x = S$ *and* $S \to y \in P$,
- *or* $x = y_1X_1 \dots y_nX_n, y = y_1x_1 \dots y_nx_n$, *where* $y_i \in T^*, x_i \in T^*N \cup T^*, X_i \in N_i$, *and* $X_i \to x_i \in P$, $1 \le i \le n$.

${}_{par}\mathfrak{R}(i) = \{L \mid L = L(G)$, *where* $G = (N_1, N_2, \dots, N_n, T, R, S)$ *is a i-PRLG*$\}$.

Theorem 2.2 (Wood hierarchy). *For all* $i \ge 1$, ${}_{par}\mathfrak{R}(i) \subset {}_{par}\mathfrak{R}(i+1)$.

Proof. See [Wood–73]. □

For more information about n-parallel right-linear grammars, see [Wood–73].

Lemma 2.1. *Let* $i \ge 1$. ${}_{dis}\mathfrak{R}(REG, i) \subseteq {}_{par}\mathfrak{R}(i)$. *That is,*
for every $n-{}_{dis}RLG$ $G = (N_1, \dots, N_n, T, R_1, \dots, R_n, S)$ *regulated by regular language* Ξ *there exists equivalent* $n-PRLG$ $G' = (N'_1, \dots, N'_n, T', S', P')$ *such that* $L(G) = L(G')$.

Proof. Let $\Xi = L(G_\Xi), G_\Xi = (N_\Xi, T_\Xi, R_\Xi, S_\Xi)$. Let $R = R_1 \cup R_2 \cup \dots \cup R_n$. We will define grammar $G' = (N'_1, \dots, N'_n, T', S', P')$ this way:

- $N'_1 = \{[A,X] \mid A \in N_1, X \in N_\Xi\}$,
- $N'_i = N_i, 2 \le i \le n$,
- $T' = T$,
- $P'_1 = \{([A_1,X], A_2, \dots, A_n) \to (t[B_1,Y], A_2, \dots, A_n) \mid$
  $\mid A_i \in N_i, 1 \le i \le n,\ X,Y \in N_\Xi$ and $f.A_1 \to tB_1 \in R_1,\ X \to fY \in R_\Xi,$
  $t \in T^*\}$,
- $P'_2 = \{([A_1,X], A_2, \dots, A_j, \dots, A_n) \to ([A_1,Y], A_2, \dots, tB_j, \dots, A_n) \mid$
  $\mid A_i \in N_i, 1 \le i \le n,\ 2 \le j \le n,\ X,Y \in N_\Xi$ and
  $f.A_j \to tB_j \in R_j,\ X \to fY \in R_\Xi,\ t \in T^*\}$,
- $P' = P'_1 \cup P'_2 \cup \{S' \to [X_1, S_\Xi]X_2 \dots X_n \mid S = X_1 \dots X_n \in G,\ X_i \in N_i,\ 1 \le i \le n\}$.

Note that P'_1 is a special case of P'_2 with $j = 1$.

Let $L_n(G) = \{x \mid S \Rightarrow^n x$ in $G,\ x \in \{N \cup T\}^*\}$ and $L_n(G') = \{x \mid S' \Rightarrow^{n+1} x$ in $G',\ x \in \{N' \cup T'\}^*\}$. We will prove that $L_n(G) = h(L_n(G'))$ for every $n \ge 0$, where h is surjective function $h : \{N'_1 \cup \dots \cup N'_n \cup T'\} \to \{N_1 \cup \dots \cup N_n \cup T\}$ defined as

$$h(w) = \begin{cases} A, \text{ if } w \in N'_1, w = [A,Y], \\ w, \text{ otherwise.} \end{cases}$$

First we will prove that $L_n(G) \subseteq h(L_n(G'))$ by induction on n:
Let $n = 0$. $L_0(G) = \{X_1X_2 \dots X_n\}$, $L_0(G') = \{[X_1,Y]X_2 \dots X_n\}$ because $S' \to [X_1,Y]X_2 \dots X_n \in P'$ and, therefore, $h(L_0(G')) = \{X_1X_2 \dots X_n\} = L_0(G)$.
Let us suppose that the claim holds for all $n \leq k$, where k is a non-negative integer.
Let $n = k+1$. Consider $w \in L_{k+1}(G)$ and a derivation $S \Rightarrow^k v \Rightarrow w$ in G, so that $v \Rightarrow w\ [p]$, where $v = C_1C_2 \dots C_{i-1}XC_{i+1} \dots C_n$, $w = C_1C_2 \dots C_{i-1}tYC_{i+1} \dots C_n, C_j \in N_j \cup \{T\}^*, 1 \leq j \leq n, p.X \to tY \in R, A \to pB \in R_\Xi$. From the induction step, $v \in h(L_k(G'))$.
As $([C_1,A]C_2 \dots C_{i-1}XC_{i+1} \dots C_n) \to ([C_1,B]C_2 \dots C_{i-1}tYC_{i+1} \dots C_n) \in P'$, $w \in h(L_{k+1}(G'))$.

Now we prove that $L_n(G) \supseteq h(L_n(G'))$ by induction on $n \geq 0$:
Let $n = 0$. By analogy with the previous part of this proof.
Let us suppose that our claim holds for all $n \leq k$, where k is a non-negative integer.
Let $n = k+1$. Consider $w \in L_{k+1}(G')$ and a derivation $S \Rightarrow^k v \Rightarrow w$ in G', where $v = [C_1,A]C_2 \dots C_{i-1}XC_{i+1} \dots C_n$, $w = [C_1,B]C_2 \dots C_{i-1}tYC_{i+1} \dots C_n, C_j \in N_j \cup \{T\}^*, 1 \leq j \leq n$. From the induction step, $h(v) \in L_k(G)$. Since $p.X \to tY \in R, A \to pB \in R_\Xi$, we have $h(w) \in L_{k+1}(G)$. □

Lemma 2.2. *Let $i \geq 1$. ${}_{dis}\mathfrak{R}(REG,i) \supseteq {}_{par}\mathfrak{R}(i)$ That is,*
for every n-PRLG $G' = (N'_1, \dots, N'_n, T', S', P')$ there exists equivalent $n-{}_{dis}RLG$ $G = (N_1, \dots, N_n, T, R_1, \dots, R_n, S)$ regulated by regular language Ξ such that $L(G) = L(G')$.

Proof. G is defined in this way

- $N_i = N'_i, 1 \leq i \leq n$;
- $T = T'$;
- $S = S'$;
- $R_i = \{r_{ij}.A_i \to t_iB_i \mid$ for the jth rule $(A_1, \dots, A_i, \dots, A_n) \to (t_1B_1, \dots, t_iB_i, \dots, t_nB_n) \in P', t_i \in T^*,\ 1 \leq j \leq |P'|\},\ 1 \leq i \leq n$.

and $\Xi = L(G_\Xi), G_\Xi = (N_\Xi, T_\Xi, R_\Xi, S_\Xi)$ is defined

- $N_\Xi = \{Q\} \cup \{Q_{ij} \mid 1 \leq i \leq n-1,\ 1 \leq j \leq |P'|\}$;
- $T_\Xi = \{r_{ij} \mid 1 \leq i \leq n,\ 1 \leq j \leq |P'|\}$;
- $R_\Xi = \{Q \to r_{1j}Q_{1j} \mid 1 \leq j \leq |P'|\} \cup$
  $\cup \{Q_{ij} \to r_{i+1j}Q_{i+1j} \mid 1 \leq i \leq n-2,\ 1 \leq j \leq |P'|\} \cup$
  $\cup \{Q_{n-1j} \to r_{nj}Q \mid 1 \leq j \leq |P'|\}$;
- $S_\Xi = Q$.

□

Theorem 2.3. *For all $i \geq 1$, ${}_{dis}\mathfrak{R}(REG,i) = {}_{par}\mathfrak{R}(i)$.*

Proof. This theorem directly follows from Lemma 2.1 and Lemma 2.2 □

The main result of this paper follows next.

Theorem 2.4. *For all* $i \geq 1$, $\Re(REG, i) \subset \Re(REG, i+1)$.

Proof. This theorem follows from Theorems 2.1, 2.2 and 2.3. □

3 CONCLUSION

As we show in Theorem 2.4 right-linear grammars with start string of length n regulated by regular language form an infinite hierarchy of languages according to the length of start string. This result is surprising because all classic language families generate the same language even if we start derivation from a start string.

REFERENCES

[Das–89] Dassow, J., Păun, Gh.: *Regulated Rewriting in Formal Language Theory*, Springer, 1989.

[Gre–69] Greibach, S. A.: An Infinite Hierarchy of Context-Free Languages, *Journal of the ACM*, Volume 16, Pages 91-106, 1969.

[Roz–73] Rozenberg, G. and Saloma, A. (eds.): *Handbook of Formal Languages*, Volumes 1 through 3, Springer, 1997.

[Rych–08] Rychnovsky, L.: *PhD Thesis*, 2008.

[Sal–73] Salomaa, A.: Formal Languages, Academic Press, New York, 1973.

[Wood–73] Wood, D.: Properties of n-Parallel Finite State Languages, Utilitas Mathematica, Winnipeg, Canada. Vol. 4, 1973. Pages 103-113.