Chapter 3 Regular Languages and Regular Grammars

何宗易

Tsung-Yi Ho
tyho@cs.nctu.edu.tw
http://idea.cs.nctu.edu.tw
Department of Computer Science
National Chiao Tung University
Hsinchu, Taiwan



Outline

1

Regular Expressions (RE)

2

Connection Between REs and Regular Languages

3

Regular Grammars

Specifying Language

How do we specify languages?

- If language is finite, you can list all of its strings.
 - $-L = \{a, aa, aba, aca\}$
- Descriptive:
 - $L = \{x \mid n_a(x) = n_b(x)\}$
- Using basic Language operations
 - $L= \{aa, ab\}^* \cup \{b\}\{bb\}^*$
- Regular languages are described using the last method

Regular Expressions

Regular expressions describe regular languages and the notation involves a combination of:

- Strings of symbols from some alphabet Σ
- Parentheses ()
- Operators +, ·, *

Regular Expressions

Important thing to remember

- A regular expression is not a language
- A regular expression is used to describe a language.
- It is incorrect to say that for a language L,
 L = (a + b + c)*
- But it's okay to say that L is described by (a + b + c)*

Regular Expressions

All finite languages can be described by regular expressions

Example:
$$(a+b\cdot c)* \longleftrightarrow \{\{a\} \cup \{bc\}\}^*$$

describes the language

$${a,bc}* = {\lambda,a,bc,aa,abc,bca,...}$$

Definition 3.1

Let Σ be a given alphabet. Then

- 1. ϕ , λ , and a ϵ Σ are all regular expressions. These are called primitive regular expressions.
- 2. If r_1 and r_2 are regular expressions, so are r_1+r_2 , $r_1\cdot r_2$, r_1^* and (r_1) .
- 3. A string is a regular expression *iff* it can be derived from the primitive regular expressions by a finite number of applications of the rules in (2).

A regular expression: $(a+b\cdot c)*\cdot(c+\varnothing)$

```
Not a regular expression: (a + b +)

a<sup>n</sup>

a<sup>+</sup>
```

Languages of Regular Expressions

L(r): language of regular expression r

Example

$$L((a+b\cdot c)^*) = \{\lambda, a, bc, aa, abc, bca, \ldots\}$$

Definition

For primitive regular expressions:

$$L(\varnothing) = \varnothing \tag{1}$$

$$L(\lambda) = \{\lambda\} \tag{2}$$

$$L(a) = \{a\} \tag{3}$$

Definition (continued)

For regular expressions r_1 and r_2

$$L(r_1 + r_2) = L(r_1) \cup L(r_2)$$
 (4)

$$L(r_1 \cdot r_2) = L(r_1) L(r_2)$$
 (5)

$$L(r_1 *) = (L(r_1))*$$
 (6)

$$L((r_1)) = L(r_1) \tag{7}$$

Regular expression: $(a+b) \cdot a^*$

$$L((a+b) \cdot a^*) = L((a+b)) L(a^*)$$

$$= L(a+b) L(a^*)$$

$$= (L(a) \cup L(b)) (L(a))^*$$

$$= (\{a\} \cup \{b\}) (\{a\})^*$$

$$= \{a,b\} \{\lambda,a,aa,aaa,...\}$$

$$= \{a,aa,aaa,...,b,ba,baa,...\}$$

Priority of Operators

• Regular expression: $r = a \cdot b + c$ $r_1 = a \cdot b$ $r_2 = c$ or $r_1 = a$ $r_2 = b + c$ $L(r) = \{ab, c\} \neq \{ab, ac\}$

 Star closure (*) precedes concatenation (·) precedes union (+)

$$\Sigma = \{a,b\}$$

• Regular expression r = (a+b)*(a+bb)

Stands for any string of a's and b's

$$L(r) = \{a,bb,aa,abb,ba,bbb,...\}$$

L(r) is the set of all strings on {a,b}, terminated by either an a or a bb

• Regular expression r = (aa)*(bb)*b

$$L(r) = \{a^{2n}b^{2m+1}: n, m \ge 0\}$$

L(r) is the set of all strings with an even number of a's followed by an odd number of b's

• For $\Sigma = \{0, 1\}$, give a regular expression r such that

 $L(r) = \{ w \in \Sigma^* : w \text{ has at least one pair of consecutive } 0 \}$

00

L(r) = { all strings with no pairs of consecutive 0s }

Regular expression

$$r = (1+01)*(0+\lambda)$$

$$r = (1*011*)*$$
Add 1 immediately after a 0 String ending in 0 String with all 1's

There are an unlimited number of REs for any given language!

Equivalent Regular Expressions

Definition:

Regular expressions r_1 and r_2

are equivalent if $L(r_1) = L(r_2)$

Example

L = { all strings without two consecutive 0 }

$$r_1 = (1+01)*(0+\lambda)$$

$$r_2 = (1*011*)*(0+\lambda)+1*(0+\lambda)$$

$$L(r_1) = L(r_2) = L$$

 r_1 and r_2 are equivalent regular expressions

- $L_1 = \{a, aa, aba, aca\}$
- $L_1 = \{a\} \cup \{aa\} \cup \{aba\} \cup \{aca\}$
- Regular expression describing L₁:

```
(a + aa + aba + aca)
```

- $L_2 = \{x \in \{0,1\}^* \mid |x| \text{ is even}\}$
- $L_2 = \{00, 01, 10, 11\}^*$
- Regular expressions describing L₂:

$$(00 + 01 + 10 + 11)^*$$

 $((0 + 1)(0 + 1))^*$

- L₃ = {x ∈ {0,1}* | x does not end in 01 }
 If x does not end in 01, then either
 x ends in 00, 10, or 11
- A regular expression that describes L₃ is:
 (0 + 1)*(00 + 10 + 11)

- L₄ = {x ∈ {0,1}* | x contains an odd number of 0s }
 Express x = yz
 y is a string of the form y=1ⁱ01^j
 In z, there must be an even number of 0's
 z = (01^k01^m)*
- A regular expression that describes L₄ is:
 (1*01*)(01*01*)*

Short Quiz

- Give regular expressions for the following language on Σ= {a, b, c}.
 - All strings containing exactly one a

Outline

1

Regular Expressions (RE)

2

Connection Between REs and Regular Languages

3

Regular Grammars

Theorem

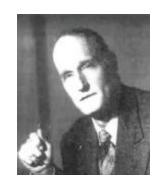
Languages
Described by
Regular Expressions

Regular Expressions

For every regular language there is a regular expression For every regular expression there is a regular language

Kleene Theorem:

Regular expressions and Finite Automata are equivalent (w.r.t. the languages they describe/accept)



Theorem - Part 1

 Languages

 Described by

 Regular Expressions

 Languages

1. For any regular expression r the language L(r) is regular

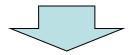
■Theorem 3.1

Theorem - Part 2

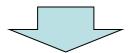
2. For any regular language L there is a regular expression r with L(r) = L

Theorem 3.2

1. For any regular expression r the language L(r) is regular



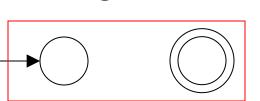
If we have any regular expression r, we can construct an NFA(DFA) that accepts L(r)



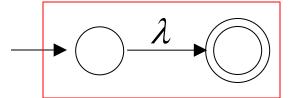
Proof by induction on the size of *r*

Induction Basis

• Primitive Regular Expressions: \emptyset , λ , α



$$L(M_1) = \emptyset = L(\emptyset)$$



$$L(M_2) = \{\lambda\} = L(\lambda)$$

regular languages

$$L(M_3) = \{a\} = L(a)$$

Inductive Hypothesis

Assume for regular expressions r_1 and r_2 that $L(r_1)$ and $L(r_2)$ are regular languages

Inductive Step

: REs are derived from these four rules:

$$L(r_1+r_2)$$

$$L(r_1 \cdot r_2)$$

We will prove:

$$L(r_1*)$$

Are regular Languages

$$L((r_1))$$

By definition of regular expressions:

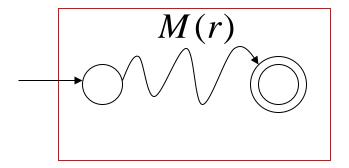
$$L(r_1 + r_2) = L(r_1) \cup L(r_2)$$

$$L(r_1 \cdot r_2) = L(r_1) L(r_2)$$

$$L(r_1 *) = (L(r_1))*$$

$$L((r_1)) = L(r_1)$$

Schematic representation of an NFA (M(r)) accepting L(r)



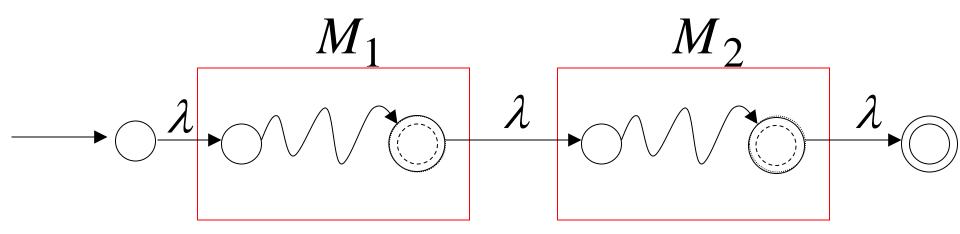
We can claim that for every NFA there is only one final state (by exercise 7, section 2.3)

Union

• NFA for $L(r_1 + r_2)$

Concatenation

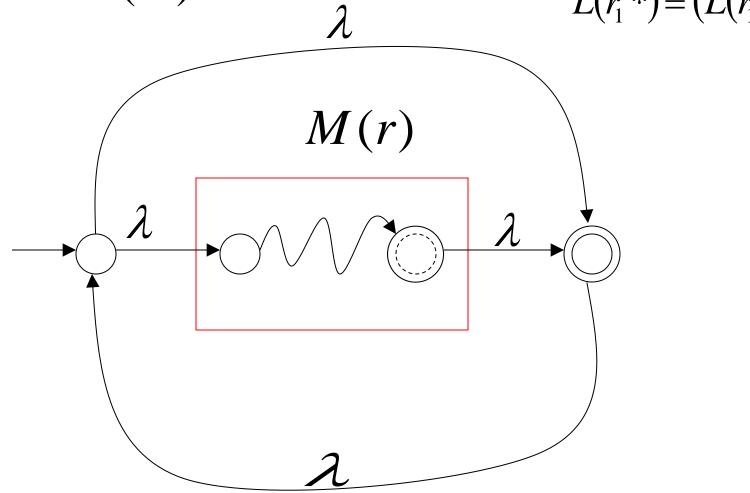
• NFA for $L(r_1r_2)$



Star Operation

• NFA for $L(r^*)$

$$L(r_1 *) = (L(r_1))*$$



By inductive hypothesis we know:

 $L(r_1)$ and $L(r_2)$ are regular languages

We also know:

Regular languages are closed under:

Union
$$L(r_1) \cup L(r_2)$$

Concatenation $L(r_1) L(r_2)$
Star $(L(r_1))*$

Therefore:

$$L(r_1 + r_2) = L(r_1) \cup L(r_2)$$

$$L(r_1 \cdot r_2) = L(r_1) L(r_2)$$

$$L(r_1 *) = (L(r_1))*$$

Are regular languages

And trivially:

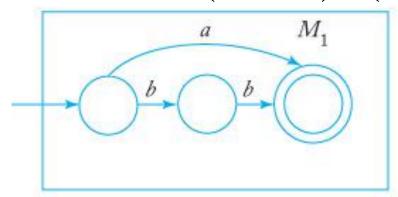
 $L((r_1))$ is a regular language

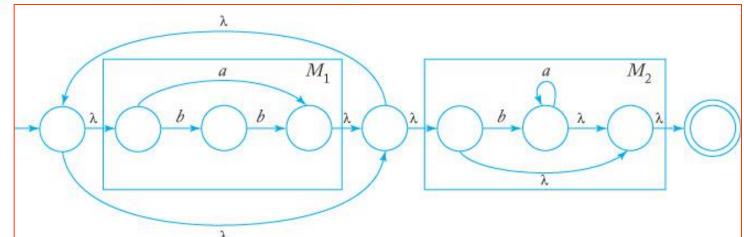
 \therefore For any regular expression r the language L(r) is regular

Example 3.7

Find an NFA that accepts L(r), where

$$r = (a+bb)*(ba*+\lambda)$$





2015)

2. For any regular language L there is a regular expression r with L(r) = L



Since any regular language has an associated NFA and hence a transition graph, all we need to do is to find a regular expression

capable of generating the labels of all the walks from q_0 to any final state.



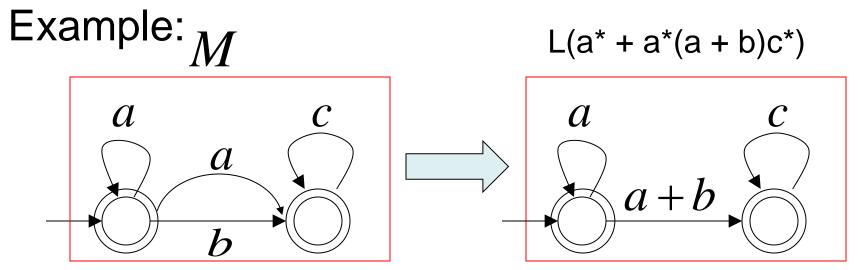
Proof by construction of regular expression

Generalized Transition Graphs (GTG)

From *M* construct the equivalent

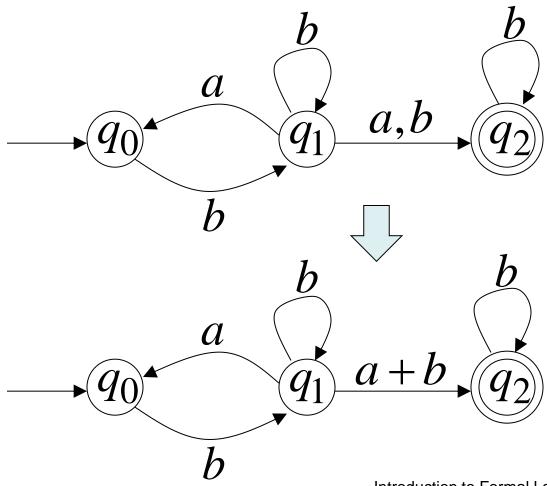
Generalized Transition Graph

in which transition labels are regular expressions



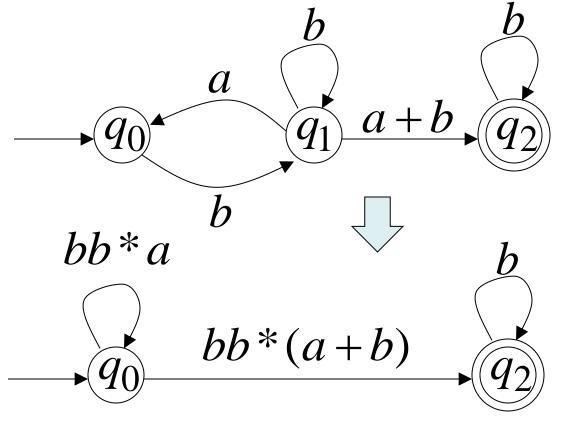
Example 3.8

GTG may have many states Enumerating all walks is time-consuming



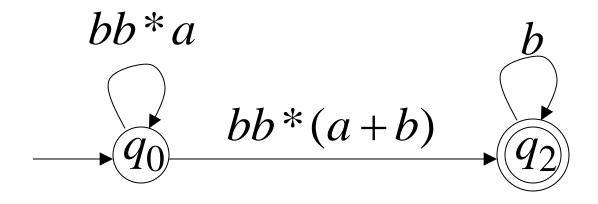
Reducing the states:

Ex. reduce q₁



Simple two-state GTG

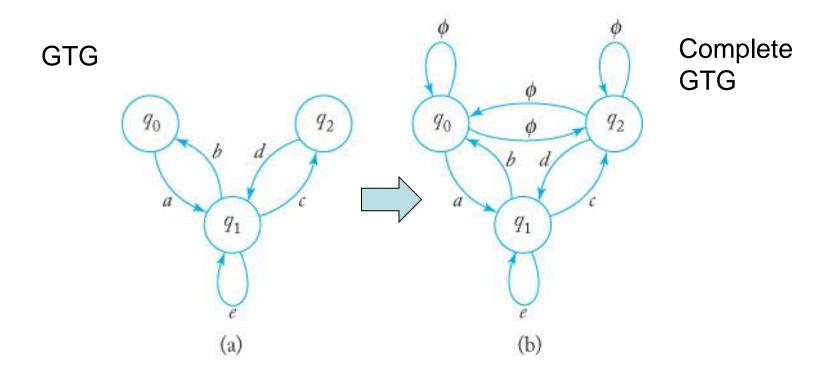
Resulting Regular Expression:



$$r = (bb*a)*bb*(a+b)b*$$

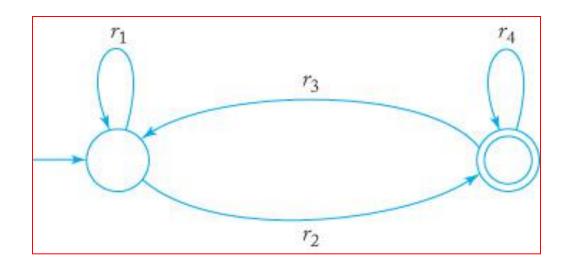
$$L(r) = L(M) = L$$

Complete GTG



- If a GTG, after conversion from an NFA, has some edges missing, we put them in and label them with φ
- ■A complete GTG with |V| vertices has exactly |V|² edges

Example 3.9



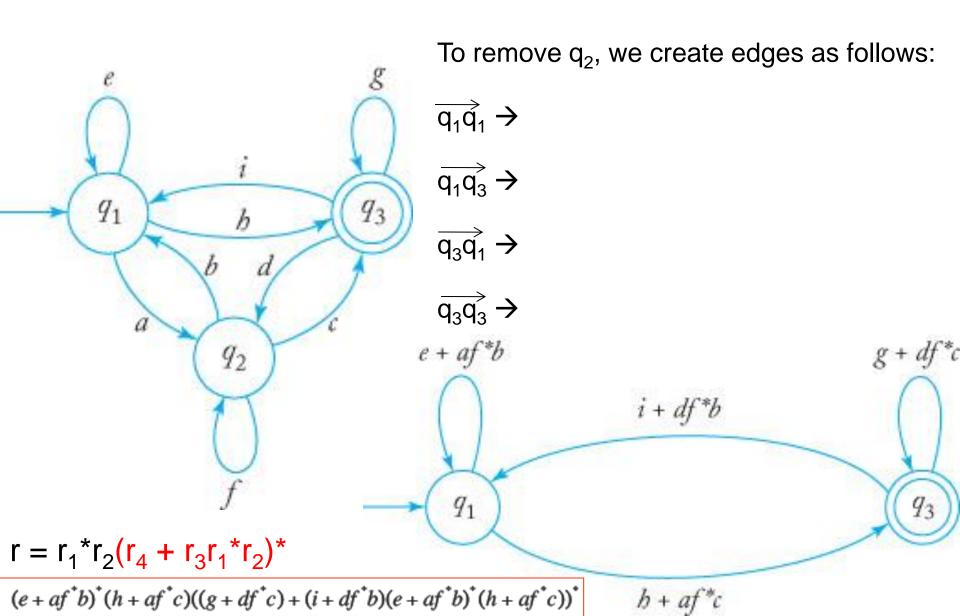
RE?

$$r = r_1 r_2 (r_4 + r_3 r_1 r_2)^*$$

How about a GTG with more than two states?

We can find an equivalent graph by removing one state at a time

Example 3.10



$NFA \rightarrow RE$

- - 1. Convert the NFA (with single final state) into a complete GTG. Let r_{ii} stand for the label of the edge from q_i to q_i.
 - 2. If the GTG has only two states with $q_i \in q_0$ and $q_i \in F$, as its associated RE is:

$$r = r_{ii}^* r_{ij} (r_{jj} + r_{ji} r_{ii}^* r_{ij})^*$$

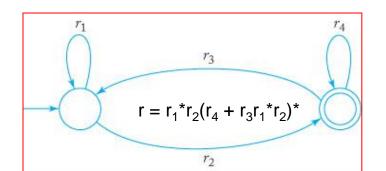
3. If the GTG has three states with $q_i \in q_0$, $q_i \in F$, and $q_k \in Q$, introduce new edges, labeled:

$$r_{pq} + r_{pk}r_{kk}r_{kq}$$

for p = i, j, q = i, j. When this is done, remove vertex q_k and its associated edges.

i + df*b

h + af *c



$NFA \rightarrow RE$

4. If the GTG has four or more states, pick a state q_k to be removed. Apply rule 3 for all pairs of states (q_i, q_j), i ≠ k, j ≠ k. At each step apply the simplifying rules

$$r + \varphi = r$$
, $r \cdot \varphi = \varphi$, $\varphi^* = \lambda$

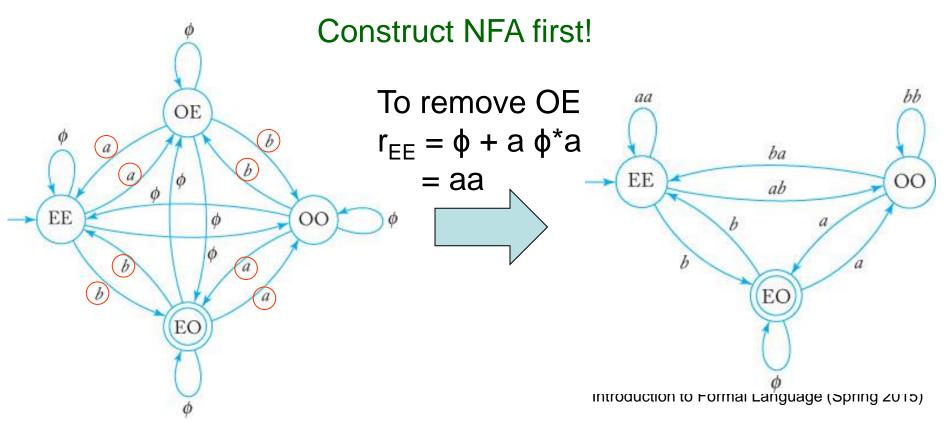
wherever possible. When this is done, remove state q_k .

5. Repeat step 2 to 4 until the correct RE is obtained.

Example 3.11

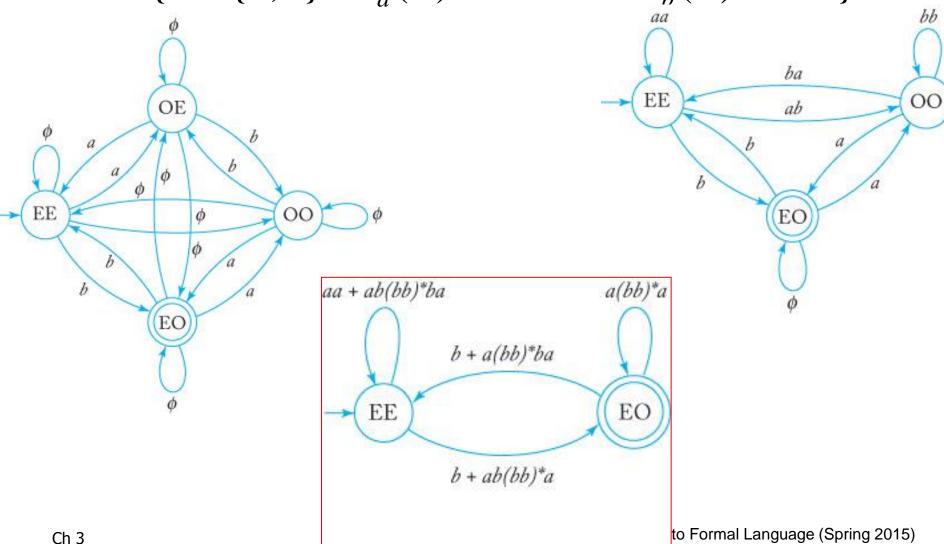
Find a RE for the language

 $L = \{w \in \{a,b\}^* : n_a(w) \text{ is even and } n_b(w) \text{ is odd}\}.$



Example 3.11

 $L = \{w \in \{a,b\}^* : n_a(w) \text{ is even and } n_b(w) \text{ is odd}\}.$



Outline

1

Regular Expressions (RE)

2

Connection Between REs and Regular Languages

3

Regular Grammars

Grammar Recap

A grammar G is defined as a 4-tuple:

$$G = (V, T, S, P)$$

where

- V is a finite set of variables
- T is a finite set of terminals
- S ∈ V, called start variable
- P is a finite set of production rules

Grammar Recap

Let G = (V, T, S, P) be a grammar. Then the set
 L(G) = {w ∈ T*: S ⇒ w}
 is the language generated by G

If w ∈ L(G), then the sequence

$$S \Rightarrow W_1 \Rightarrow W_2 \Rightarrow ... \Rightarrow W_n \Rightarrow W$$

is a derivation of the sentence w.

• S, w₁, w₂, ..., w_n are called sentential forms

Linear Grammars

Grammars with at most one variable at the right side of a production

Examples:
$$S \rightarrow aSb$$

$$S \rightarrow aSb$$

$$S \to \lambda$$

$$S \rightarrow Ab$$

$$A \rightarrow aAb$$

$$A \rightarrow \lambda$$

Another Linear Grammar

Grammar
$$G: S \to A$$

$$A \to aB \mid \lambda$$

$$B \to Ab$$

$$L(G) = \{a^n b^n : n \ge 0\}$$

A Non-Linear Grammar

Grammar $G: S \rightarrow SS$

$$S \to \lambda$$

$$S \rightarrow aSb$$

$$S \rightarrow bSa$$

$$L(G) = \{w: n_a(w) = n_b(w)\}$$

Number of a in string w

Right-Linear Grammars

• All productions have form: $A \rightarrow xB$

or
$$A \rightarrow x$$

• Example: $S \rightarrow abS$ $S \rightarrow a$

string of terminals

Left-Linear Grammars

• All productions have form: $A \rightarrow Bx$

or
$$A \rightarrow x$$

• Example: $S \rightarrow Aab$

$$S \rightarrow Aab$$

$$A \rightarrow Aab \mid B$$

$$B \rightarrow a$$

string of terminals

Regular Grammars

A regular grammar is either right-linear or left-linear grammar

Examples:

$$G_1$$
 O

$$S \rightarrow abS$$

$$S \rightarrow a$$

$$G_2$$

$$S \rightarrow Aab$$

$$A \rightarrow Aab \mid B$$

$$B \rightarrow a$$

$$G_3$$

$$S \to A$$

$$A \rightarrow aB \mid \lambda$$

$$B \rightarrow Ab$$

Introduction to Formal Language (Spring 2015)

Observation

Regular grammars generate regular languages

A regular grammar is always linear, but not all linear grammars are regular.

G₃ is linear grammar but not regular grammar

$$G_3$$
 $S \to A$
 $A \to aB \mid \lambda$
 $B \to Ab$

Example 3.13

Regular grammars generate regular languages

$$G_1$$
 $S \rightarrow Aab$ $S \rightarrow abS$ $A \rightarrow Aab \mid B$ $S \rightarrow a$ $B \rightarrow a$

$$L(G_1) = (ab) * a$$

$$L(G_2) = aab(ab) *$$

Theorem

Languages
Generated by
Regular Grammars

Regular Grammars

Theorem - Part 1

Languages
Generated by
Regular Grammars

Regular Languages

Any regular grammar generates a regular language

■Theorem 3.3

Theorem - Part 2

 Languages

 Generated by

 Regular Grammars

 Languages

Any regular language is generated by a regular grammar

■Theorem 3.4

Proof – Part 1

The language L(G) generated by any regular grammar G is regular

The case of Right-Linear Grammars

Let G be a right-linear grammar

We will prove: L(G) is regular

Proof idea: We will construct NFA M with L(M) = L(G)

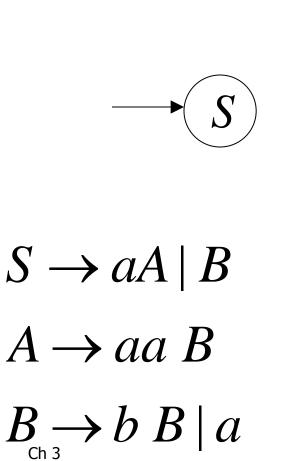
Grammar G is right-linear

Example: $S \rightarrow aA \mid B$

 $A \rightarrow aa B$

 $B \rightarrow b B \mid a$

Construct NFA M such that every state is a grammar variable:

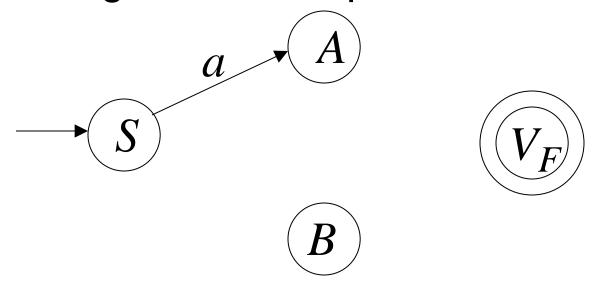




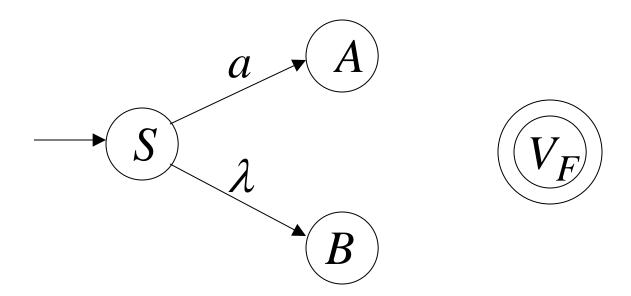




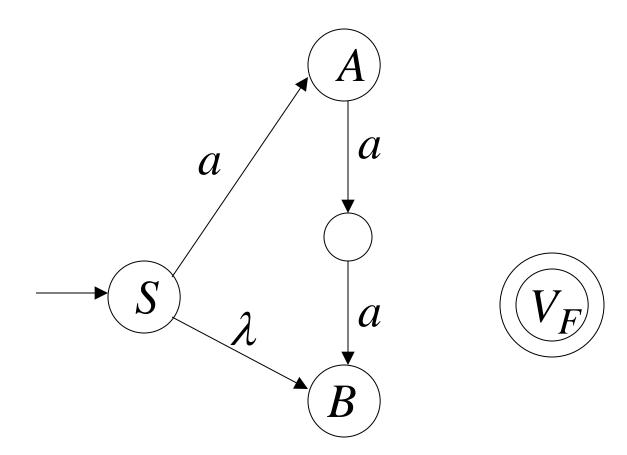
Add edges for each production:



$$S \rightarrow aA$$

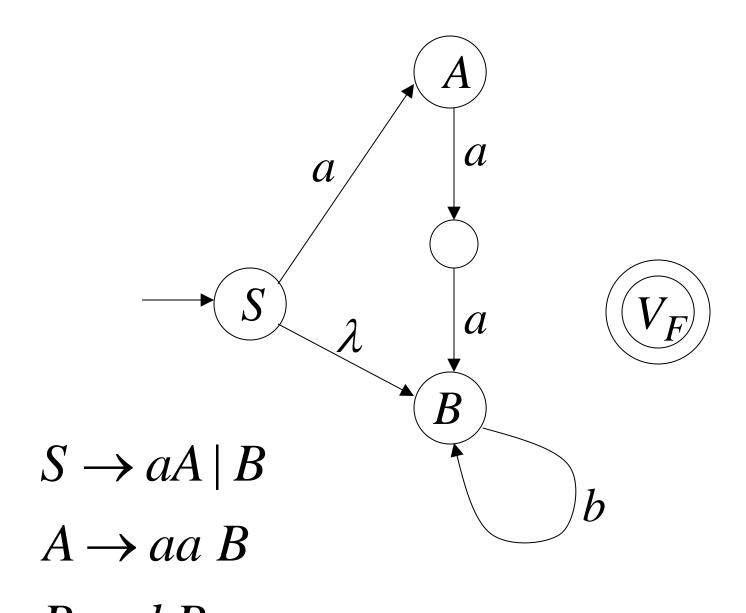


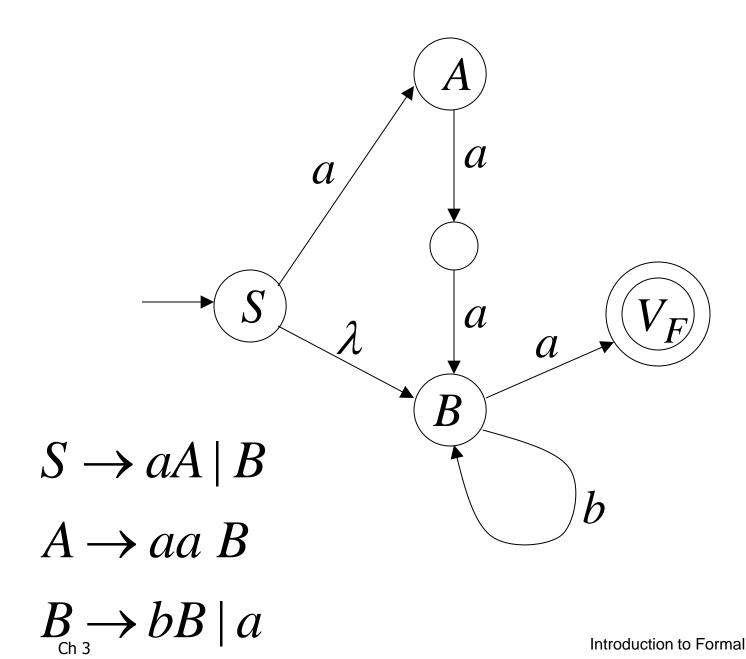
$$S \rightarrow aA \mid B$$

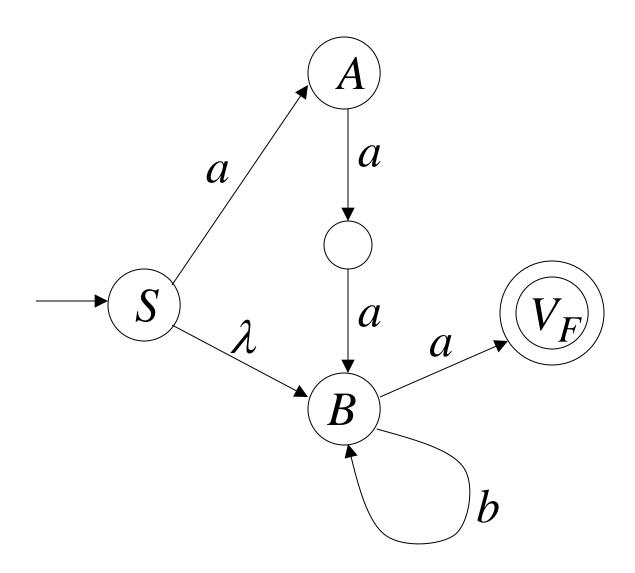


$$S \rightarrow aA \mid B$$
 $A \rightarrow aa \mid B$

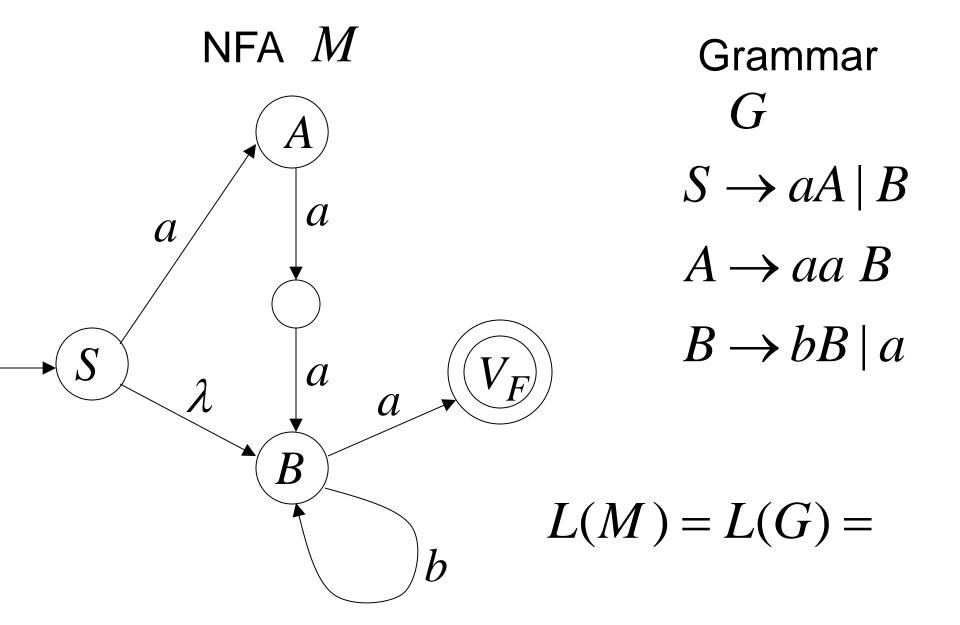
$$A \rightarrow aa B$$







 $S \Rightarrow aA \Rightarrow aaaB \Rightarrow aaabB \Rightarrow aaaba$



In General

A right-linear grammar G

has variables: V_0, V_1, V_2, \dots

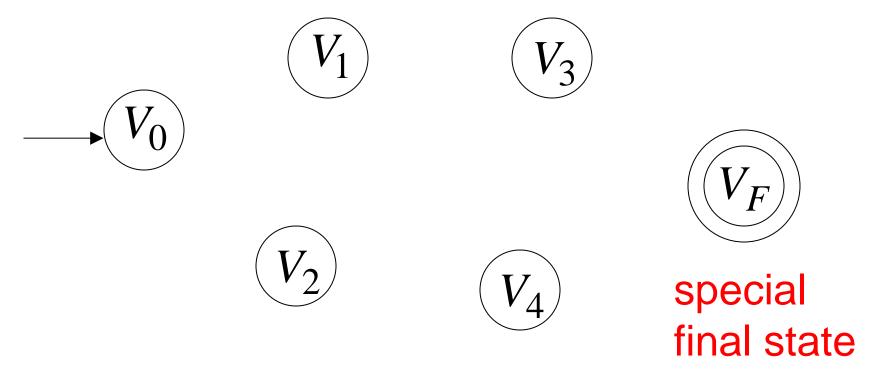
and productions:
$$V_i \rightarrow a_1 a_2 \cdots a_m V_j$$

or

$$V_i \rightarrow a_1 a_2 \cdots a_m$$

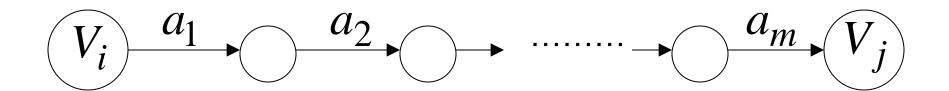
We construct the NFA M such that:

each variable V_i corresponds to a node:



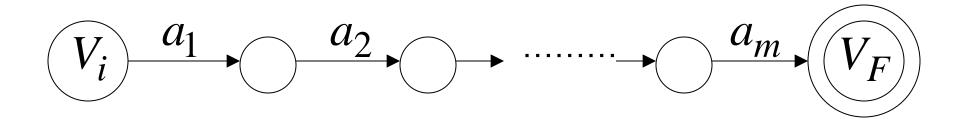
For each production: $V_i \rightarrow a_1 a_2 \cdots a_m V_j$

we add transitions and intermediate nodes



For each production: $V_i \rightarrow a_1 a_2 \cdots a_m$

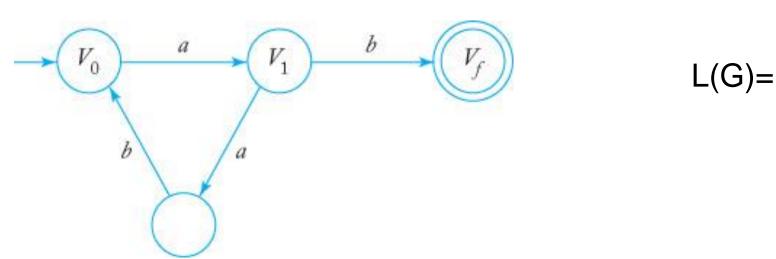
we add transitions and intermediate nodes



Example 3.15

 Construct a FA that accepts the language generated by the grammar

$$V_0 \rightarrow aV_1$$
,
 $V_1 \rightarrow abV_0|b$



The case of Left-Linear Grammars

Let G be a left-linear grammar

We will prove: L(G) is regular

Proof idea:

We will construct a right-linear grammar G' with $L(G) = L(G')^R$

Since G is left-linear grammar the productions look like:

$$A \rightarrow Ba_1a_2\cdots a_k$$

$$A \rightarrow a_1 a_2 \cdots a_k$$

Construct right-linear grammar G'

Left G

$$A \rightarrow Ba_1a_2\cdots a_k$$

$$A \rightarrow Bv$$



Right G^{\prime}

$$A \rightarrow a_k \cdots a_2 a_1 B$$

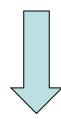
$$A \rightarrow v^R B$$

• Construct right-linear grammar G'

Left _{linear} *G*

$$A \rightarrow a_1 a_2 \cdots a_k$$

$$A \rightarrow v$$



Right linear G'

$$A \rightarrow a_k \cdots a_2 a_1$$

$$A \rightarrow v^R$$

It is easy to see that: $L(G) = L(G')^R$

Since G' is right-linear, we have:

$$L(G')$$
 \longrightarrow $L(G)^R$ Regular Regular Language Language

Proof - Part 2

Example 1 Languages | Compare |

Generated by Regular Grammars | Compare |

Regular Grammars | C

Any regular language L is generated by some regular grammar G

Any regular language L is generated by some regular grammar G

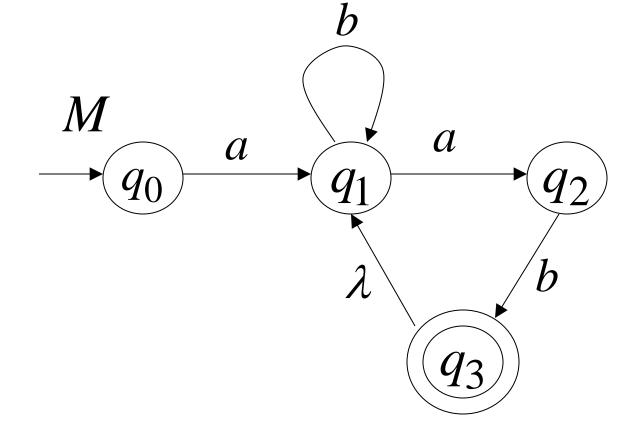
Proof idea:

Let M be the NFA with L = L(M)

Construct from M to a regular grammar Gsuch that L(M) = L(G)

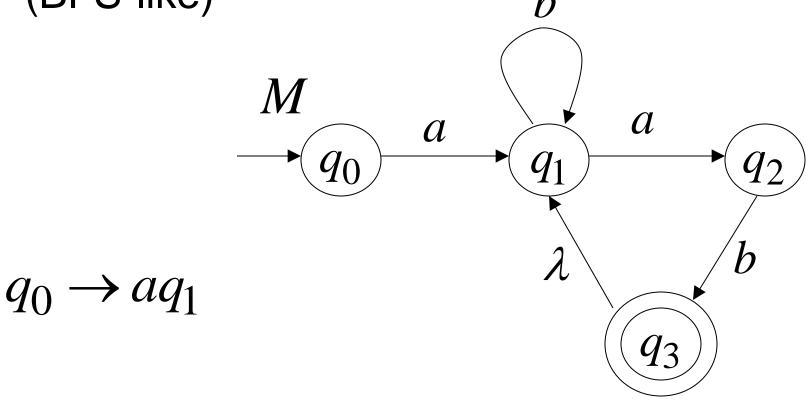
Since L is regular there is an NFA M such that L = L(M)

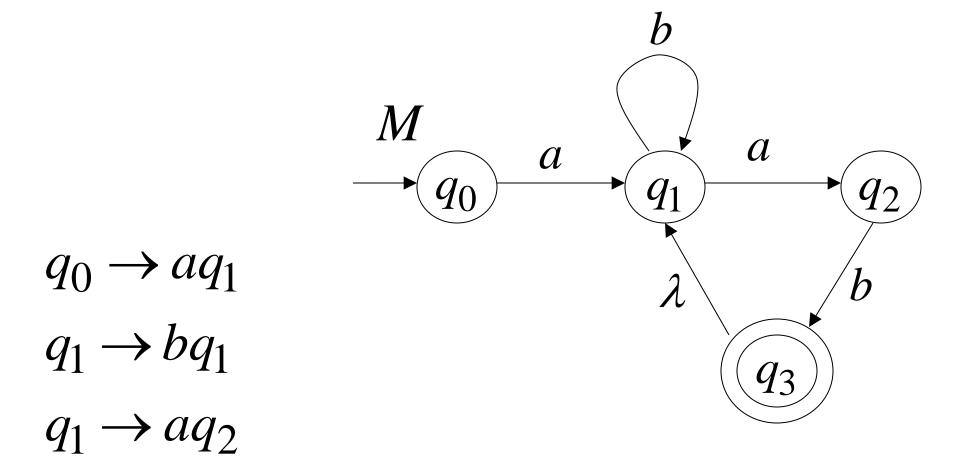
Example:

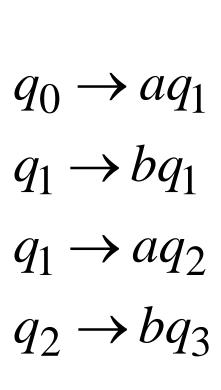


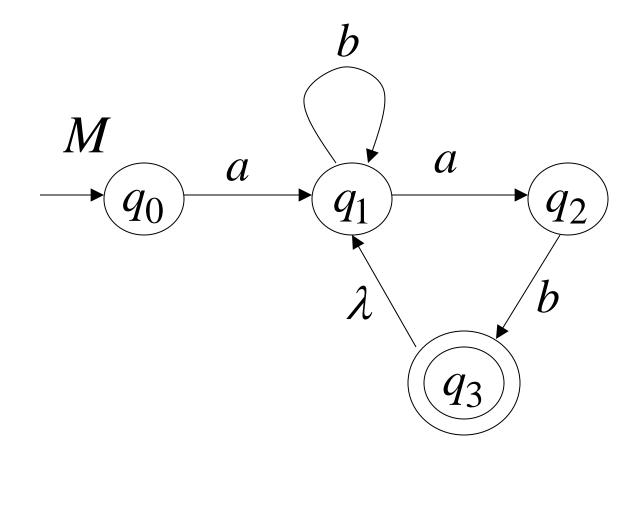
$$L = L(M)$$

Convert M to a right-linear grammar (BFS-like)

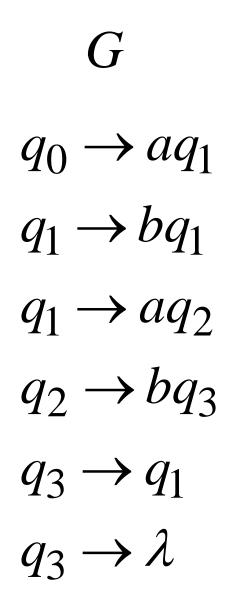


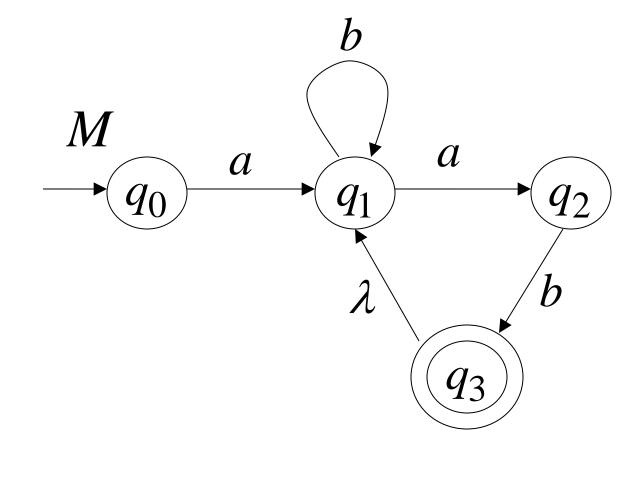






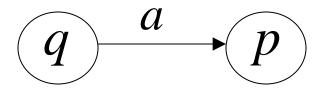
$$L(G) = L(M) = L$$

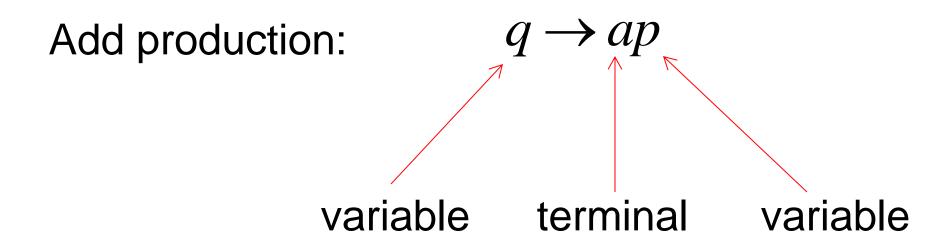




In General

For any transition:





For any final state:

$$(q_f)$$

Add production:

$$q_f \to \lambda$$

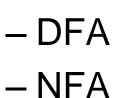
Since G is right-linear grammar

G is also a regular grammar

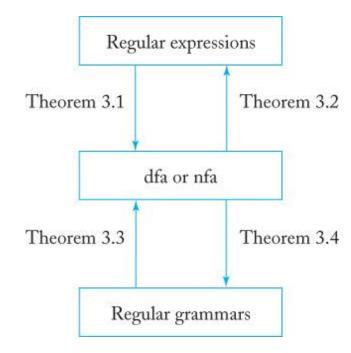
with
$$L(G) = L(M) = L$$

Summary

 We now have several ways of describing regular languages:



- RE
- -RG



Questions?