

# Einführung in die Computerlinguistik – reguläre Sprachen und endliche Automaten

Dozentin: Wiebke Petersen

03.11.2009

# Formal language

## Definition

Eine **formale Sprache**  $L$  ist eine Menge von Wörtern über einem Alphabet  $\Sigma$ , also  $L \subseteq \Sigma^*$ .

## Definition

Ein **Wort** ist eine endliche Kette/Folge  $x_1 \dots x_n$  von Symbolen/Zeichen eines Alphabets ( $n \geq 0$ ). Das Wort, das aus null Zeichen besteht heißt **leeres Wort** und wird mit  $\epsilon$  bezeichnet.

## Definition

Ein **Alphabet**  $\Sigma$  ist eine nichtleere endliche Menge von **Symbolen / Zeichen**.

# Describing formal languages by enumerating all words

- Peter says that Mary has fallen off the tree.
- Oskar says that Peter says that Mary has fallen off the tree.
- Lisa says that Oskar says that Peter says that Mary has fallen off the tree.
- ...

# Describing formal languages by enumerating all words

- Peter says that Mary has fallen off the tree.
- Oskar says that Peter says that Mary has fallen off the tree.
- Lisa says that Oskar says that Peter says that Mary has fallen off the tree.
- ...

The set of strings of a natural language is infinite.

The enumeration does not gather generalizations.

# Describing formal languages by grammars

## Grammar

- A formal grammar is a **generating device** which can generate (and analyze) strings/words.
- Grammars are finite rule systems.
- The set of all strings generated by a grammar is the formal language generated by the grammar.

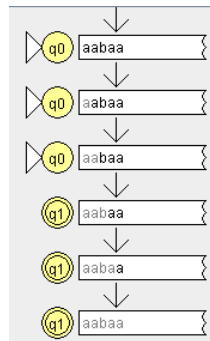
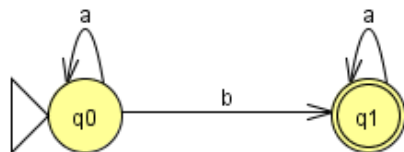
$$\begin{array}{l}
 S \rightarrow NP VP \quad VP \rightarrow V \quad NP \rightarrow D N \\
 D \rightarrow \text{the} \quad N \rightarrow \text{cat} \quad V \rightarrow \text{sleeps}
 \end{array}$$

Generates: the cat sleeps

# Describing formal languages by automata

## Automaton

- An automaton is a **recognizing device** which accepts strings/words.
- The set of all strings accepted by an automaton is the formal language accepted by the automaton.



# Sprachbeschreibung

## Zusammenhang nach Klabunde 1998

“Formale Sprachen besitzen strukturelle Eigenschaften.  
Grammatiken sind Erzeugungssysteme für formale Sprachen.  
Automaten sind Erkennungssysteme für formale Sprachen.”

**Vorsicht:** per Definition besitzen formale Sprachen keine strukturellen Eigenschaften; uns interessieren aber nur solche mit strukturellen Eigenschaften, die von einer Grammatik erzeugt werden können. Außerdem können Grammatiken auch für die Analyse (Erkennung) formaler Sprachen und endliche Automaten für ihre Erzeugung genutzt werden.

# Regular expressions

## RE: syntax

The set of **regular expressions**  $RE_{\Sigma}$  over an alphabet  $\Sigma = \{a_1, \dots, a_n\}$  is defined by:

- $\emptyset$  is a regular expression.
- $\epsilon$  is a regular expression.
- $a_1, \dots, a_n$  are regular expressions
- If  $a$  and  $b$  are regular expressions over  $\Sigma$  then
  - $(a + b)$
  - $(a \bullet b)$
  - $(a^*)$

are regular expressions too.

(The brackets are frequently omitted w.r.t. the following dominance scheme:  
\* dominates • dominates +)



# Regular expressions

## RE: semantics

Each regular expression  $r$  over an alphabet  $\Sigma$  describes a formal language  $L(r) \subseteq \Sigma^*$ .

**Regular languages** are those formal languages which can be described by a regular expression.

The function  $L$  is defined inductively:

- $L(\emptyset) = \emptyset$ ,  $L(\epsilon) = \{\epsilon\}$ ,  $L(a_i) = \{a_i\}$
- $L(a + b) = L(a) \cup L(b)$
- $L(a \bullet b) = L(a) \circ L(b)$
- $L(a^*) = L(a)^*$

# Aufgaben für Übungssitzung (1)

## Exercise 1

*Find a regular expression which describes the regular language  $L$  (be careful: at least one language is not regular!)*

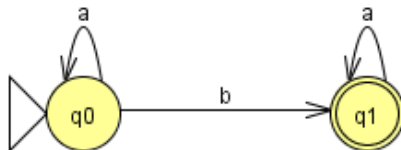
- *$L$  is the language over the alphabet  $\{a, b\}$  with  $L = \{aba, \epsilon, aa, bbb\}$ .*
- *$L$  is the language over the alphabet  $\{a, b\}$  which consists of all words which start with a nonempty string of  $b$ 's followed by at least one  $a$  followed by any number of  $b$ 's*
- *$L$  is the language over the alphabet  $\{a, b\}$  such that every  $a$  has a  $b$  immediately to its left.*
- *$L$  is the language over the alphabet  $\{a, b\}$  which consists of all words which contain an uneven number of  $a$ 's.*
- *$L$  is the language of all palindromes over the alphabet  $\{a, b\}$ .*

# Deterministic finite-state automaton (DFSA)

## Definition

A *deterministic finite-state automaton* is a tuple  $\langle Q, \Sigma, \delta, q_0, F \rangle$  with:

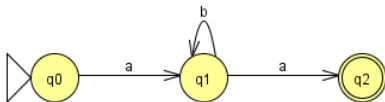
- 1 a finite, non-empty set of *states*  $Q$
- 2 an alphabet  $\Sigma$  with  $Q \cap \Sigma = \emptyset$
- 3 a partial *transition* function  $\delta : Q \times \Sigma \rightarrow Q$
- 4 an *initial state*  $q_0 \in Q$  and
- 5 a set of *final/accept states*  $F \subseteq Q$ .



accepts:  $L(a^*ba^*)$

# partial/total transition function

## FSA with partial transition function



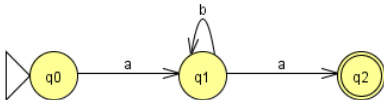
accepts  $ab^*a$

	a	b
q0	q1	
q1	q2	q1
q2		

transition table

# partial/total transition function

FSA with partial transition function

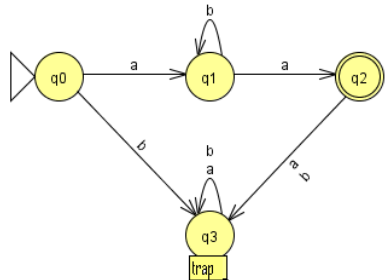


accepts  $ab^*a$

	a	b
q0	q1	
q1	q2	q1
q2		

transition table

FSA with complete transition function



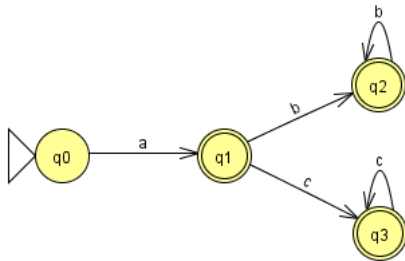
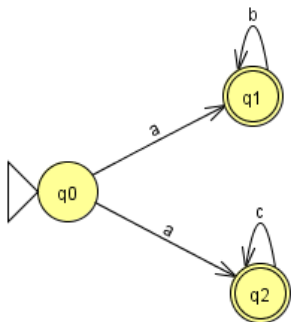
accepts  $ab^*a$

	a	b
q0	q1	q3
q1	q2	q1
q2	q3	q3
q3	q3	q3

transition table

# Example DfSA / NDFSA

The language  $L(ab^* + ac^*)$  is accepted by



# Nondeterministic finite-state automaton NDFSA

## Definition

A *nondeterministic finite-state automaton* is a tuple  $\langle Q, \Sigma, \Delta, q_0, F \rangle$  with:

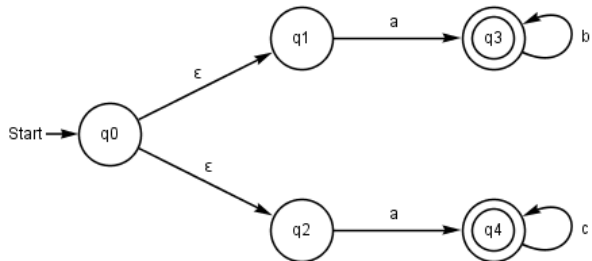
- 1 a finite non-empty set of *states*  $Q$
- 2 an alphabet  $\Sigma$  with  $Q \cap \Sigma = \emptyset$
- 3 a **transition relation**  $\Delta \subseteq Q \times \Sigma \times Q$
- 4 an *initial state*  $q_0 \in Q$  and
- 5 a set of *final states*  $F \subseteq Q$ .

## Theorem

A language  $L$  can be accepted by a DFSA iff  $L$  can be accepted by a NFSA.

Note: Even automata with  $\epsilon$ -transitions accept the same languages like NDFSA's.

# Automaton with $\epsilon$ -transition





# Aufgaben für Übungssitzung (2)

## Exercise 2

Give an FSA for each of the following languages over the alphabet  $\{a, b\}$  (and try to make it deterministic):

- $L = \{w \mid \text{between each two 'a's in } w \text{ there are at least three 'b's}\}$
- $L = \{w \mid w \text{ is any word except "bab"}\}$
- $L = \{w \mid w \text{ does not contain the infix "ab"}\}$
- $L = \{w \mid w \text{ contains at most two 'a's}\}$
- $L = \{w \mid w \text{ contains an uneven number of 'a's}\}$
- $L((b^*a)^*ab^*)$
- $L(a^*(ab)^*)$
- $L(aa^*b)$ .
- $L((bb^* + ab^*a))$

# Hausaufgaben zum 12.11.2009

Üben Sie bitte die folgenden Aufgabentypen (in der Sitzung findet eine kurze Überprüfung statt):

- Formulierung eines regulären Ausdrucks für eine verbal beschriebene reguläre Sprache
- Zeichnen eines endlichen Automaten, der die reguläre Sprache erkennt, die von einem regulären Ausdruck beschrieben wird
- Umwandlung eines endlichen Automaten mit  $\epsilon$ -Übergängen in einen ohne  $\epsilon$ -Übergänge.

Sie können zum Üben das Programm `exorciser` benutzen und bei Problemen und Fragen die CL Tutoren fragen.