Κεφάλαιο 2:
Τυπικές γλώσσες

Τυπικές γλώσσες (i)

Βασικές έννοιες

- Αλφάβητο Σ
- Σύμβολο a
- Συμβολοσειρά α
- Μήκος συμβολοσειράς $|\alpha|$
- Σύνολο συμβολοσειρών μήκους n Σ^n
- Σύνολο όλων των συμβολοσειρών

$$\Sigma^* = \bigcup_{n=0}^{\infty} \Sigma^n$$

Τυπικές γλώσσες (ii)

Βασικές έννοιες *(συνέχεια)*

- Κενή συμβολοσειρά ϵ
- Παράθεση συμβολοσειρών $\alpha\beta$
- Παράθεση συμβολοσειράς με τον εαυτό της

$$\begin{aligned} \alpha^0 &= \epsilon \\ \alpha^{n+1} &= \alpha\alpha^n \end{aligned}$$

- Πρόθεμα, επίθεμα, υποσυμβολοσειρά

Τυπικές γλώσσες (iii)

Βασικές έννοιες *(συνέχεια)*

- Γλώσσα $L \subseteq \Sigma^*$
- Ένωση γλωσσών

$$L_1 \cup L_2 = \{ \alpha \mid \alpha \in L_1 \vee \alpha \in L_2 \}$$

- Παράθεση γλωσσών

$$L_1 L_2 = \{ \alpha\beta \mid \alpha \in L_1 \wedge \beta \in L_2 \}$$

- Παράθεση γλώσσας με τον εαυτό της

$$\begin{aligned} L^0 &= \{ \epsilon \} \\ L^{n+1} &= LL^n \end{aligned}$$

- Κλείσιμο ή άστρο του Kleene

$$L^* = \bigcup_{n=0}^{\infty} L^n \qquad L^+ = LL^*$$

Τυπικές γλώσσες (iv)

Γεννητικά μοντέλα

- Γραμματική $G = (T, N, P, S)$
  - T : τερματικά σύμβολα — a
  - N : μη τερματικά σύμβολα — A
  - P : κανόνες παραγωγής — $\alpha \to \beta$
  - S : αρχικό σύμβολο
- Παραγωγές: αν $\alpha, \beta, \gamma, \delta \in (T \cup N)^*$
  και $(\alpha \to \beta) \in P$
  τότε $\gamma\alpha\delta \Rightarrow \gamma\beta\delta$
- Γλώσσα: $L(G) = \{ \alpha \in T^* \mid S \Rightarrow^+ \alpha \}$

Τυπικές γλώσσες (ν)

Ιεραρχία Chomsky

- Τύπου 0: όλες οι γραμματικές, $\alpha \to \beta$
- Τύπου 1: γραμματικές με συμφραζόμενα (context-sensitive), $\alpha \to \beta$ με $|\alpha| \leq |\beta|$
- Τύπου 2: γραμματικές χωρίς συμφραζόμενα (context-free) $A \to \beta$
- Τύπου 3: κανονικές γραμματικές (regular) $A \to \mathrm{a}B$ ή $A \to \mathrm{a}$
- Ειδική περίπτωση: γλώσσες που παράγουν την κενή συμβολοσειρά

Τυπικές γλώσσες (vi)

Αναγνωριστές

- Τύπου 0: μηχανή Turing
- Τύπου 1: γραμμικά περιορισμένη μηχανή Turing
- Τύπου 2: αυτόματα στοίβας (push-down automata)
  - Χρήσιμα στη συντακτική ανάλυση
- Τύπου 3: πεπερασμένα αυτόματα (finite automata)
  - Χρήσιμα στη λεκτική ανάλυση

Κανονικές γλώσσες (i)

- Κανονικές γραμματικές
  - Μόνο κανόνες $A \to \mathrm{a}B$ ή $A \to \mathrm{a}$
  - ισοδύναμα $A \to B\mathrm{a}$ ή $A \to \mathrm{a}$
- Κανονικές εκφράσεις (regular expressions)
  - Κενή συμβολοσειρά: ϵ
  - Κάθε σύμβολο του Σ: a
  - Παράθεση δύο κανονικών εκφράσεων: (rs)
  - Διάζευξη δύο κανονικών εκφράσεων: $(r|s)$
  - Κλείσιμο (ή άστρο) Kleene: (r^*)
- Συντομογραφίες:
  - απαλοιφή περιττών παρενθέσεων
  - r^+ $[\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_n]$ $[\mathrm{a}_1 - \mathrm{a}_2]$ $r?$.

Κανονικές γλώσσες (ii)

Παραδείγματα κανονικών εκφράσεων

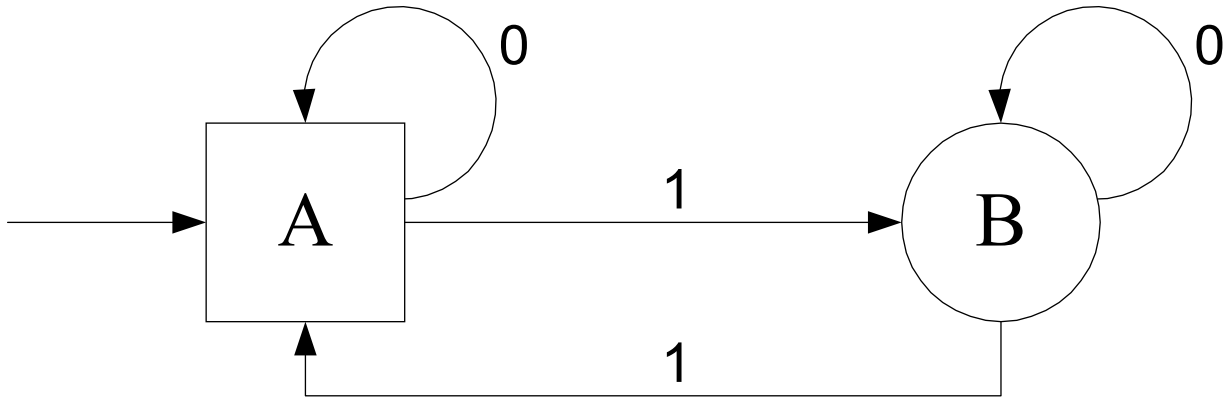
- Ακέραιες σταθερές χωρίς πρόσημο στην Pascal
  - ένα ή περισσότερα δεκαδικά ψηφία

$$[0\text{-}9]^{+}$$

- Αριθμητικές σταθερές χωρίς πρόσημο στη C
  - ακέραιο μέρος που δεν αρχίζει με μηδέν, εκτός αν είναι μηδενικό (γιατί;)
  - προαιρετικά: υποδιαστολή και κλασματικό μέρος
  - προαιρετικά: εκθέτης με ή χωρίς πρόσημο

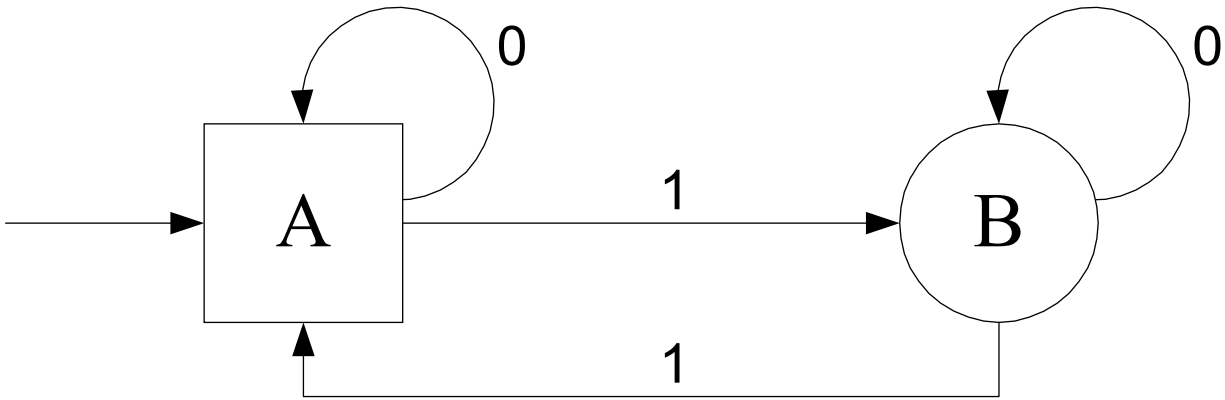
$$([1\text{-}9][0\text{-}9]^{*}|0)(\backslash.[0\text{-}9]^{+})?((\mathrm{E}|\mathrm{e})(+|-)?[0\text{-}9]^{+})?$$

Πεπερασμένα αυτόματα (i)



- Καταστάσεις και μεταβάσεις
- Ντετερμινιστικά (ΝΠΑ), μη ντετερμινιστικά (ΜΠΑ) και ΜΠΑ με κενές μεταβάσεις (ΜΠΑ-ϵ)
- Αρχική κατάσταση, τελικές καταστάσεις

Πεπερασμένα αυτόματα (ii)



- Ποια γλώσσα αναγνωρίζει;
- Τη γλώσσα των συμβολοσειρών που αποτελούνται από 0 και 1 και περιέχουν άρτιο αριθμό 1

Κανονικές γλώσσες, ανασκόπηση

Αναγωγές και ισοδυναμίες

- κανονική γραμματική $\Rightarrow$ ΜΠΑ-ϵ
- ΜΠΑ-ϵ $\Rightarrow$ κανονική γραμματική

- κανονική έκφραση $\Rightarrow$ ΜΠΑ-ϵ
- ΜΠΑ-ϵ $\Rightarrow$ κανονική έκφραση

- ΜΠΑ-ϵ $\Rightarrow$ ΝΠΑ
- Ελαχιστοποίηση ΝΠΑ

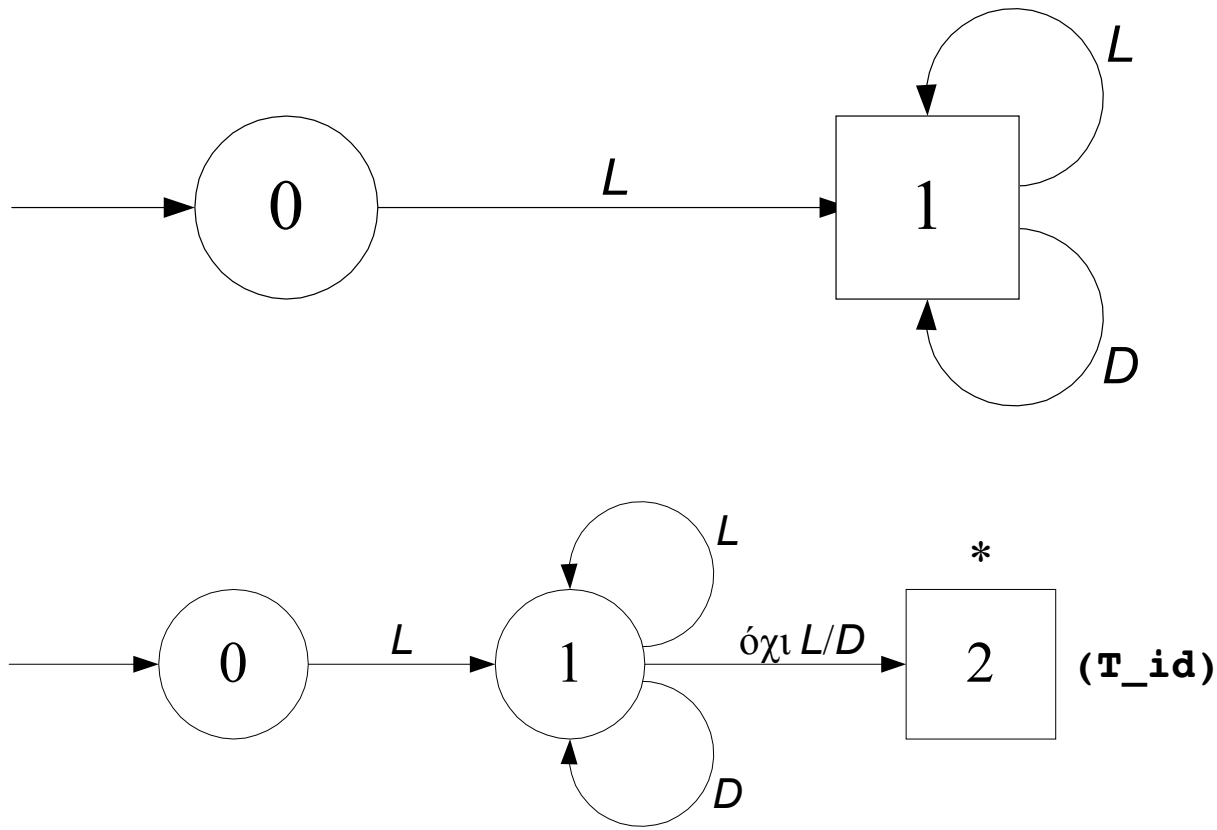
Κεφάλαιο 3:
Λεκτική ανάλυση

Λεκτική ανάλυση

- Λεκτικές μονάδες (tokens)
- Αναγνωρίζονται με πεπερασμένα αυτόματα που:
  - διαβάζουν ενδεχομένως περισσότερους χαρακτήρες
  - οπισθοδρομούν αν χρειαστεί
  - διαθέτουν έξοδο που χρησιμοποιείται στη συντακτική ανάλυση
- Ειδικός συμβολισμός: διαγράμματα μετάβασης

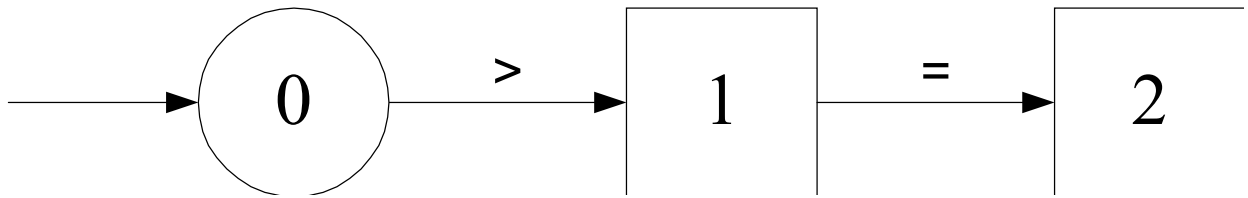
Διαγράμματα μετάβασης (i)

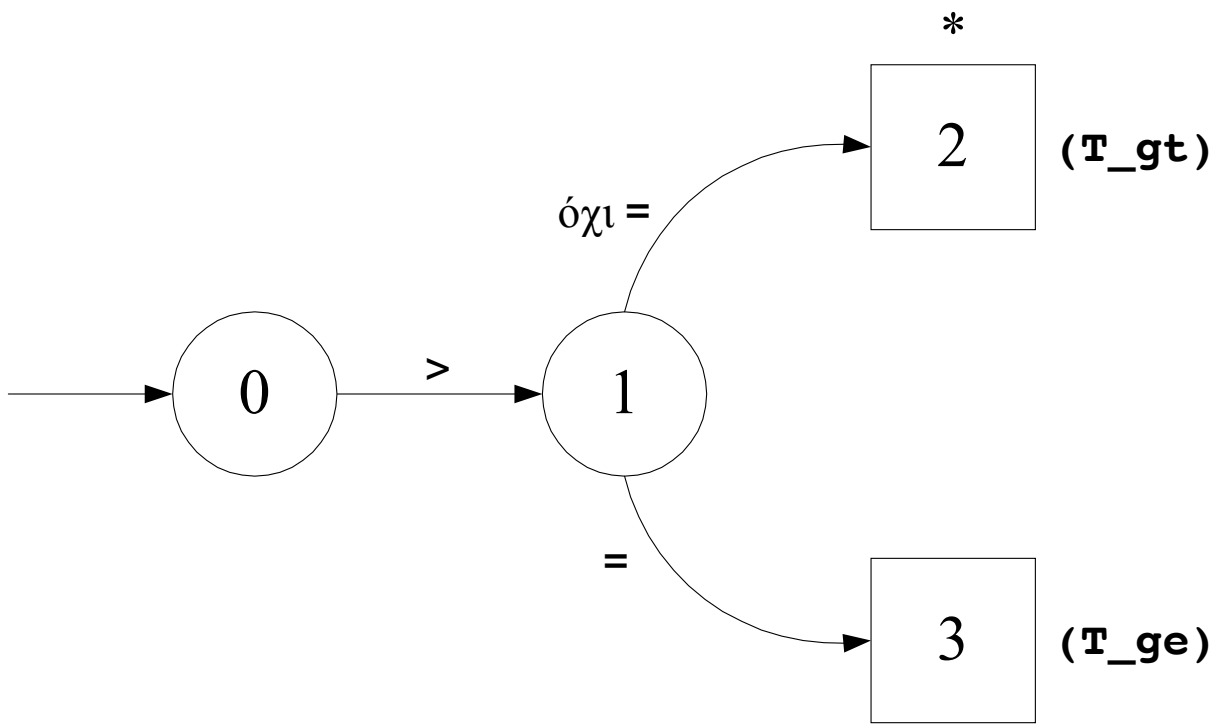
- Αναγνωριστικά της Pascal



Διαγράμματα μετάβασης (ii)

- Τελεστές > και >=





Κατασκευή του ΛΑ (i)

- Καταγραφή και ταξινόμηση χαρακτήρων

$$mapping \quad : \quad (\text{ASCII} \cup \{\,\texttt{EOF}\,\}) \rightarrow \Sigma$$

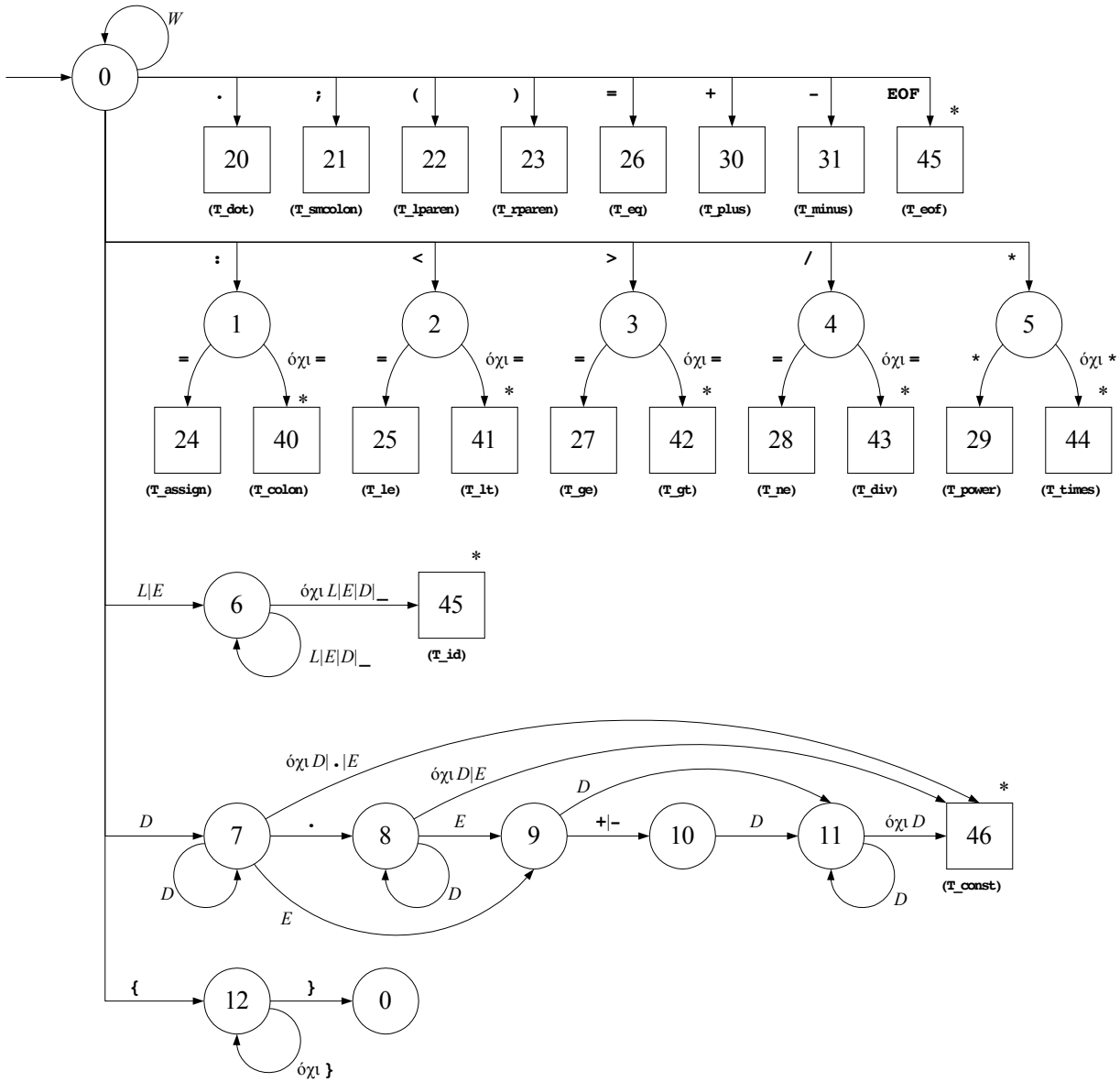
- Καταγραφή και ταξινόμηση λεκτικών μονάδων
  - Κωδικοποίηση λεκτικών μονάδων
  - Ακολουθία χαρακτήρων (lexeme)
- Σχεδίαση του διαγράμματος μετάβασης
- Υλοποίηση του λεκτικού αναλυτή

Κατασκευή του ΛΑ (ii)

- Επιμέρους θέματα
  - Τρόπος διαχωρισμού λεκτικών μονάδων
  - Σχόλια
  - Διάκριση πεζών / κεφαλαίων γραμμάτων
  - Ενδιάμεση μνήμη (buffer)
  - Ανάνηψη από σφάλματα

Κατασκευή του ΛΑ (iii)

Σχεδίαση συνολικού διαγράμματος μετάβασης



Κατασκευή του ΛΑ (iv)

- Εναλλακτικοί τρόποι υλοποίησης:
  - Χειρωνακτικά
  - Με πίνακα μεταβάσεων
  - Με το μεταεργαλείο flex