Μεταγλωττιστές

Γιώργος Δημητρίου

Μάθημα 2ο

Αλφάβητα και Γλώσσες

- Αλφάβητο: Ένα μη κενό και πεπερασμένο σύνολο συμβόλων
- Γλώσσα: Ένα οποιοδήποτε υποσύνολο των συμβολοσειρών ενός αλφαβήτου (οι προτάσεις της γλώσσας, πχ. προγράμματα)
  - Πεπερασμένη ή μη πεπερασμένη
  - Ποιες προτάσεις ανήκουν σε μια γλώσσα;
  - Υπολογίσιμες γλώσσες

Παραδείγματα

- Αλφάβητο: {‘0’,‘1’}
  Γλώσσα: {“1”,“110”,“01”}
- Αλφάβητο: {‘0’,‘1’}
  Γλώσσα: $\{0^n 1^n, n>0\}$
- Αλφάβητο: λεκτικές μονάδες C
  Γλώσσα: προγράμματα C
- Αλφάβητο: ελληνικές λέξεις
  Γλώσσα: φυσική ελληνική

Βασικές Έννοιες

- Η κενή συμβολοσειρά: ε
- Η κενή γλώσσα: $\varnothing$
- Παράθεση γλωσσών:

  $LM = \{ xy \mid x \in L \text{ και } y \in M \}$

  $\varnothing L = L\varnothing = \varnothing$, $L\{\varepsilon\} = \{\varepsilon\}L = L$

  $L^{n+1} = LL^n$, $L^0 = \{\varepsilon\}$
- Ένωση γλωσσών:

  $L \cup M = \{ x \mid x \in L \text{ ή } x \in M \}$

  $\varnothing \cup L = L \cup \varnothing = L$

Κλείσιμο Kleene

Κλείσιμο Kleene:

$$L^* = L^0 \cup L^1 \cup L^2 \cup \ldots$$

- $\varepsilon \in L^*$

Θετικό κλείσιμο Kleene:

$$L^+ = LL^* = L^1 \cup L^2 \cup \ldots$$

- $\varepsilon \in L^+$ μόνο αν $\varepsilon \in L$

Γραμματικές

- $G = (V_N, V_T, P, S)$
- Μη τερματικά σύμβολα V_N
- Τερματικά σύμβολα V_T, $V_T \cap V_N = \emptyset$
- Συντακτικοί κανόνες P:

  $\alpha \to \beta$

  όπου $\alpha \in V^+$, $\beta \in V^*$, $V = V_T \cup V_N$
- Αρχικό σύμβολο $S \in V_N$

Παράδειγμα

- $V_N = \{S, A, B\}$
- $V_T = \{‘0’, ‘1’\}$
- $P = \{$

$S \rightarrow 0A \mid 1B$

$A \rightarrow A0BA \mid B1BA \mid \varepsilon$

$1B \rightarrow 01BA \mid 1$

$0B \rightarrow 1A \mid 0$

$\}$

Ιεραρχία Γραμματικών (Chomsky)

- Τύπου 0: Ελεύθερες γραμματικές

  $\alpha \rightarrow \beta$, όπου $\alpha \in V^{+}$, $\beta \in V^{*}$

- Τύπου 1: Γραμματικές με συμφραζόμενα

  $\mu A \nu \rightarrow \mu \beta \nu$, όπου $\mu,\nu \in V^{*}$, $A \in V_N$, $\beta \in V^{+}$

- Τύπου 2: Γραμματικές χωρίς συμφραζόμενα

  $A \rightarrow \beta$, όπου $A \in V_N$, $\beta \in V^{*}$

- Τύπου 3: Κανονικές γραμματικές

  $A \rightarrow b$, όπου $A \in V_N$, $b \in V_T \cup \{\varepsilon\}$

  $A \rightarrow bC$, όπου $A,C \in V_N$, $b \in V_T$

Διαδικασία Παραγωγής

- Ξεκινώντας με τη συμβολοσειρά που περιέχει μόνο το αρχικό σύμβολο της γραμματικής,
- *Παράγουμε* διαδοχικά νέες συμβολοσειρές:
  - Αντικαθιστάμε μια υποσυμβολοσειρά α της τρέχουσας συμβολοσειράς με μια άλλη β, εάν υπάρχει συντακτικός κανόνας $\alpha \rightarrow \beta$,
    - συμβολικά: αν η τρέχουσα συμβολοσειρά έχει τη μορφή xαy, όπου $x,y \in V^*$, τότε γράφουμε $x\alpha y \Rightarrow x\beta y$
- Μέχρι η τρέχουσα συμβολοσειρά να αποτελείται μόνο από τερματικά σύμβολα.

Γραμματικές και Γλώσσες

- Γλώσσα μιας γραμματικής: Η γλώσσα που παράγεται από τη γραμματική
- Ισοδύναμες γραμματικές: Αυτές που παράγουν την ίδια γλώσσα
- Τυπικές γλώσσες: Οι γλώσσες που μπορούν να παραχθούν από γραμματικές
- Τύποι γλωσσών: Οι τύποι της γραμματικής που τις παράγουν

Ανάλυση Σύνταξης

- Δεδομένης μιας γραμματικής:
- 1. Να αποφανθούμε εάν μια πρόταση που αποτελείται από τερματικά σύμβολα της γραμματικής ανήκει στη γλώσσα που περιγράφει η γραμματική
- 2. Αν ναι, να βρούμε μια παραγωγή της γραμματικής που δίνει αυτή την πρόταση

Εργαλεία Ανάλυσης Σύνταξης

- Θεωρητικά μοντέλα αφηρημένων μηχανών για την επίλυση προβλημάτων σαν αποδοχή των προτάσεων μιας γραμματικής (μηχανές Turing – αυτόματα)
- Εφαρμόζονται στους μεταγλωττιστές για τα προβλήματα της λεκτικής και της συντακτικής ανάλυσης
- Διάφορα είδη αυτομάτων για τους διαφορετικούς τύπους γραμματικών

Κανονικές Γραμματικές (ΚΓ)

- Οι λιγότερο ισχυρές γραμματικές
- Οι γλώσσες τους αναγνωρίζονται από τα πεπερασμένα αυτόματα
- Μικρότερη πολυπλοκότητα ανάλυσης σύνταξης
- Εφαρμογή στη λεκτική ανάλυση των μεταγλωττιστών

Κανονικές Εκφράσεις (ΚΕ)

- Κάθε ΚΕ r πάνω σε ένα αλφάβητο Σ περιγράφει μια κανονική γλώσσα L(r) του Σ:

  $L(\varepsilon) = \{\varepsilon\}$

  $L(a) = \{a\},\ a \in \Sigma$

  $L(rs) = L(r)L(s)$

  $L(r|s) = L(r) \cup L(s)$

  $L(r^*) = (L(r))^*,\ L(r^+) = (L(r))^+$

- Άλλοι συμβολισμοί, όπως a? [01a-z]

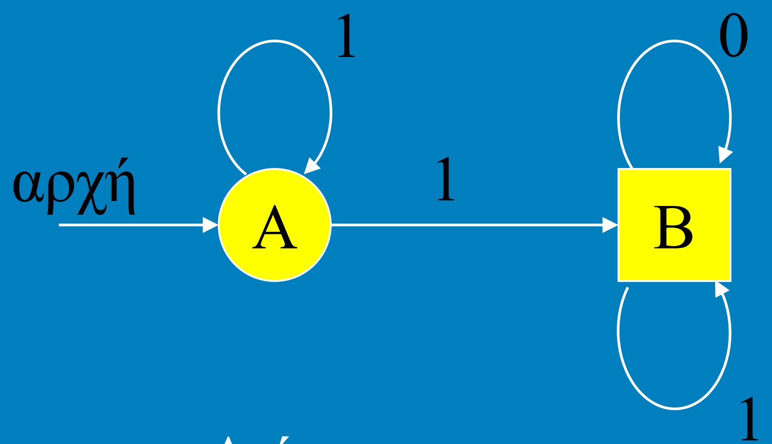
Πεπερασμένα Αυτόματα (ΠΑ)

- $M = (K, \Sigma, \delta, S, F)$
- Πεπερασμένο σύνολο καταστάσεων K
- Αλφάβητο εισόδου Σ
- Συνάρτηση μετάβασης καταστάσεων δ

  $\delta(X,y) = Z$, όπου $X,Z \in K$, $y \in \Sigma$

  – Ντετερμινιστικά / Μη ντετερμινιστικά ΠΑ
- Αρχική κατάσταση $S \in K$
- Σύνολο τελικών καταστάσεων $F \subseteq K$

Παράσταση ΠΑ



Διάγραμμα

	0	1	⊣
A		A,B	
B	B	B	Αποδοχή

Πίνακας

- Το αυτόματο που αναγνωρίζει τη γλώσσα με συμβολοσειρές τουλάχιστον ενός ‘1’, ακολουθούμενου από 0 ή περισσότερα ‘0’ ή ‘1’.

Λειτουργία ΝΠΑ

- Η κίνηση του ΝΠΑ περιγράφεται ως εξής:

1. Διάβασε το επόμενο σύμβολο από την είσοδο και προχώρησε το δείκτη εισόδου κατά μία θέση
2. Πήγαινε στην επόμενη κατάσταση με βάση την παρούσα κατάσταση και το σύμβολο που διάβασες

Λειτουργία ΜΠΑ και ΜΠΑ-ε

- Στα ΜΠΑ επιλέγεις μία από τις πιθανά πολλαπλές δυνατές κινήσεις που ορίζονται από τη συνάρτηση μετάβασης
- Στα ΜΠΑ-ε μπορείς σε κάποιες περιπτώσεις να κινηθείς χωρίς ανάγνωση συμβόλου από την είσοδο (αυθόρμητη μετάβαση στην επόμενη κατάσταση)

Αναγνώριση ή Απόρριψη

- Αν ολοκλήρωσες την ανάγνωση της εισόδου και βρίσκεσαι σε τελική κατάσταση, αναγνωρίζεις τη συμβολοσειρά, αλλιώς την απορρίπτεις
- Στα ΜΠΑ, καθώς και στα ΜΠΑ-ε, εξαντλείς όλες τις δυνατές επιλογές κίνησης, πριν απορρίψεις τη συμβολοσειρά

Μετατροπή ΜΠΑ σε ΝΠΑ

- Κατασκευή υποσυνόλων:
  - Το σύνολο καταστάσεων του ΝΠΑ περιέχει υποσύνολα του συνόλου καταστάσεων του ΜΠΑ
  - Κάθε μετάβαση πρέπει να γίνεται σε μια το πολύ κατάσταση
- Απαιτούνται μέχρι 2^n καταστάσεις του ΝΠΑ για n καταστάσεις του ΜΠΑ.

Αριθμός Καταστάσεων ΠΑ

- Σε ένα ΠΑ μπορεί να έχουμε:
  - απρόσιτες καταστάσεις
  - ισοδύναμες καταστάσεις
- Αφαιρώντας τις απρόσιτες και συμπτύσσοντας τις ισοδύναμες καταστάσεις βρίσκουμε ένα ισοδύναμο ΠΑ με ελάχιστο αριθμό καταστά-σεων

Γραμματικές χωρίς Συμφραζόμενα (ΓΧΣ)

- Πιο ισχυρές από τις ΚΓ
- Οι γλώσσες τους αναγνωρίζονται από τα αυτόματα στοίβας
- Μεγαλύτερη πολυπλοκότητα ανάλυσης σύνταξης
- Εφαρμογή στη συντακτική ανάλυση των μεταγλωττιστών

Αριστερότερη/Δεξιότερη Παραγωγή

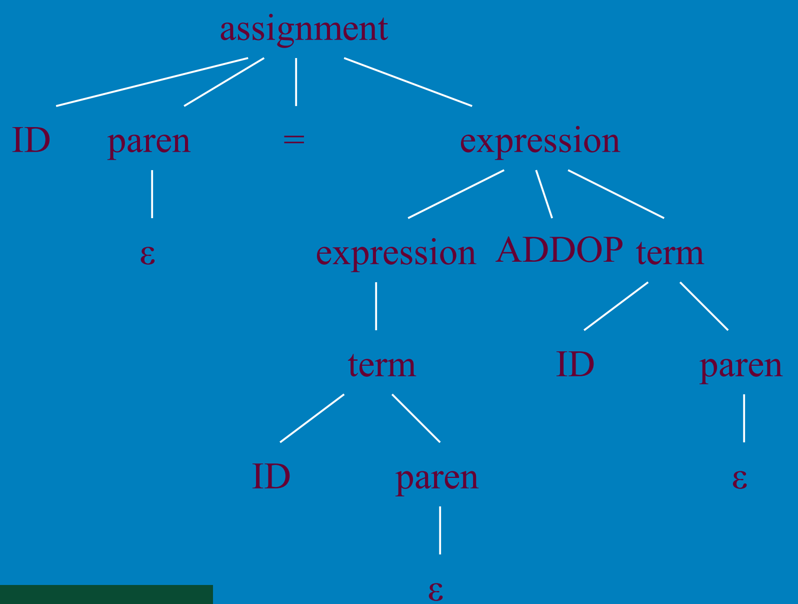
- Αν κατά τη διαδικασία παραγωγής από μια ΓΧΣ αντικαθιστούμε το αριστερότερο μη τερματικό σύμβολο, παίρνουμε μια αριστερότερη παραγωγή
  - Παράδειγμα από γραμματική 1^{ου} μαθήματος:

$$\text{assignment} \Rightarrow \text{ID paren} = \text{expression} \Rightarrow_L \text{ID} = \text{expression}$$
$$\Rightarrow \text{ID} = \text{term} \Rightarrow \text{ID} = \text{ID paren} \Rightarrow \text{ID} = \text{ID}$$

- Αντίστοιχα για δεξιότερη παραγωγή

Δέντρο Ανάλυσης Σύνταξης (ΔΣΑ)

- Απεικόνιση της διαδικασίας παραγωγής
  - Ρίζα: το αρχικό σύμβολο
  - Φύλλα: μόνο τερματικά σύμβολα
- Παράδειγμα από γραμματική 1ου μαθήματος:



Διφορούμενες Γραμματικές

- Μια ΓΧΣ ονομάζεται διφορούμενη, όταν υπάρχει πρότασή της που έχει περισσότερα από ένα ΔΑΣ, ή ισοδύναμα που παράγεται με περισσότερες από μία αριστερότερες (ή περισσότερες από μία δεξιότερες) παραγωγές
- Μια ΓΧΣ ονομάζεται *εγγενώς* διφορούμενη όταν δεν υπάρχει ισοδύναμή της μη διφορούμενη γραμματική

Παράδειγμα: Ξεκρέμαστο else

- Διφορούμενη γραμματική:

  stmt → IF cond stmt ELSE stmt | IF cond stmt | S1 | S2

  cond → C1 | C2

- Ισοδύναμη μη διφορούμενη γραμματική:

  stmt → matched | unmatched

  matched → IF cond matched ELSE matched | S1 | S2

  unmatched → IF cond stmt |
  IF cond matched ELSE unmatched

  cond → C1 | C2

- Συμβολοσειρά: IF C1 IF C2 S1 ELSE S2

Παράδειγμα: Αριθμητικές Εκφράσεις

- Διφορούμενη γραμματική:

  expr → expr + expr | expr - expr | expr * expr |
  expr / expr | ID | CONST | (expr)

- Ισοδύναμη μη διφορούμενη γραμματική:

  expr → expr + term | expr - term | term

  term → term * factor | term / factor | factor

  factor → ID | CONST | (expr)

- Συμβολοσειρά: ID + ID - CONST / ID * ID

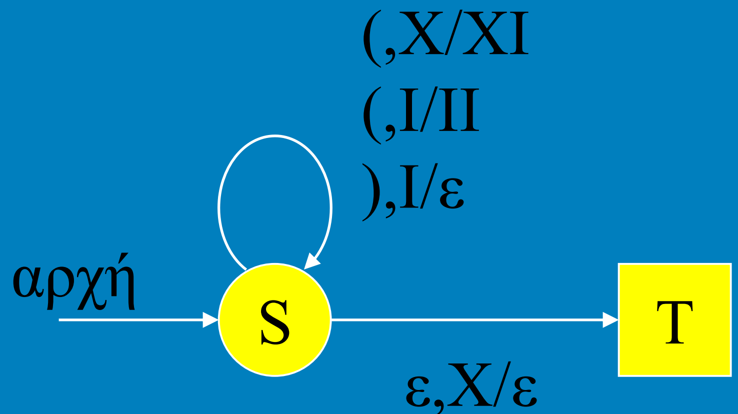
Αυτόματα Στοίβας (ΑΣ)

- $M = (K, \Sigma, \Gamma, \delta, q_0, Z_0, F)$
- Πεπερασμένο σύνολο καταστάσεων K
- Αλφάβητο εισόδου Σ
- Αλφάβητο στοίβας Γ
- Συνάρτηση μετάβασης καταστάσεων δ

  $\delta(q,X,\alpha) = (p,\gamma) \mid q,p \in K,\ X \in \Gamma,\ \alpha \in \Sigma \cup \{\varepsilon\},\ \gamma \in \Gamma^*$

  – Ντετερμινιστικά / Μη ντετερμινιστικά ΑΣ
- Αρχική κατάσταση $q_0 \in K$
- Αρχικό σύμβολο στοίβας $Z_0 \in \Gamma$
- Σύνολο τελικών καταστάσεων $F \subseteq K$

Παράσταση ΑΣ



Διάγραμμα

	X	I	⊣
S	‘(’ ⇒ S, XI ε ⇒ T,ε	‘(’ ⇒ S, II ‘)’ ⇒ S,ε	
T			Αποδοχή

Πίνακας

- Το αυτόματο που αναγνωρίζει ζευγάρια παρενθέσεων που ταιριάζουν:
  - Δύο καταστάσεις: S και T
  - Δύο σύμβολα στοίβας: X (αρχικό) και I

Λειτουργία ΑΣ

- Η κίνηση του ΑΣ περιγράφεται ως εξής:

1. Αν $\alpha \neq \varepsilon$, διάβασε το επόμενο σύμβολο από την είσοδο και προχώρησε το δείκτη εισόδου κατά μία θέση
2. Αντικατάστησε το Χ στην κορυφή της στοίβας με τα σύμβολα του γ
3. Πήγαινε στην επόμενη κατάσταση με βάση την παρούσα κατάσταση και τα σύμβολα Χ,α

Τερματισμός Λειτουργίας ΑΣ

- Η συμβολοσειρά εισόδου αναγνωρίζεται – αναλογα με την υλοποίηση – σε μία από τις ακόλουθες περιπτώσεις:
  - Όταν η συμβολοσειρά έχει ολοκληρωθεί (με το χαρακτήρα τερματισμού $) και το αυτόματο βρίσκεται σε τελική κατάσταση
  - Όταν η συμβολοσειρά έχει ολοκληρωθεί (με το χαρακτήρα τερματισμού $) και η στοίβα είναι άδεια ή περιέχει το αρχικό σύμβολο
- Σε άλλες περιπτώσεις απορρίπτεται

Ισοδυναμία Αναπαραστάσεων

- Οι ΚΓ, οι ΚΕ και τα ΠΑ είναι ισοδύναμες αναπαραστάσεις μιας κανονικής γλώσσας, και η ισοδυναμία αυτή μπορεί να αποδειχτεί με αναγωγή από τη μια αναπαράσταση στην άλλη
- Υπάρχει αντίστοιχη ισοδυναμία μεταξύ ΓΧΣ και ΑΣ