Θεωρία Υπολογισμού

Γλώσσες και Γραμματικές Χωρίς Συμφραζόμενα

2003-04

Γλώσσες και Γραμματικές Χωρίς Συμφραζόμενα

- Γλώσσες και γραμματικές χωρίς συμφραζόμενα.
- Αυτόματα στοίβας.
- Ιδιότητες των γλωσσών χωρίς συμφραζόμενα.
- Εφαρμογές στη συντακτική ανάλυση.

Γραμματικές Χωρίς Συμφραζόμενα

- Οι γραμματικές χωρίς συμφραζόμενα μελετήθηκαν αρχικά από γλωσσολόγους (για παράδειγμα τον Noam Chomsky).
- **Παράδειγμα:**

$$Sentence \longrightarrow NounPhrase\ Verb$$
$$NounPhrase \longrightarrow Adjective\ NounPhrase$$
$$NounPhrase \longrightarrow Noun$$
$$Noun \longrightarrow boy$$
$$Adjective \longrightarrow little$$
$$Verb \longrightarrow run$$

Γραμματικές Χωρίς Συμφραζόμενα

- Ο BNF (Backus-Naur-Form) συμβολισμός που χρησιμοποιούμε για να περιγράφουμε το συντακτικό γλωσσών προγραμματισμού έχει ουσιαστικά τις ίδιες έννοιες με τις γραμματικές χωρίς συμφραζόμενα.
- **Παράδειγμα:**

$$Expression ::= Expression + Expression$$
$$Expression ::= Expression * Expression$$
$$Expression ::= (Expression)$$
$$Expression ::= \mathbf{id}$$

- Ο ορισμός των παραπάνω αριθμητικών εκφράσεων δεν μπορεί να γίνει με κάποια κανονική έκφραση.

Ο Ρόλος των Γραμματικών

- Η χρήση γραμματικών προσφέρει σημαντικά πλεονεκτήματα στους σχεδιαστές γλωσσών προγραμματισμού, τους υλοποιητές μεταγλωττιστών και τους προγραμματιστές:
  - Μια γραμματική δίνει **ακριβή ορισμό** του συντακτικού μιας γλώσσας.
  - Κάποιες κλάσεις γραμματικών επιτρέπουν την αυτόματη κατασκευή αποδοτικών συντακτικών αναλυτών που αναγνωρίζουν αν ένα πρόγραμμα είναι συντακτικά σωστό (και αν δεν είναι υποδεικνύουν το λάθος που υπάρχει).
  - Η εισαγωγή μιας καινούριας έννοιας σε μια γλώσσα προγραμματισμού (και η υλοποίηση της) είναι πιο εύκολη αν υπάρχει ήδη μια γραμματική για τη γλώσσα (και οι μεταγλωττιστές είναι βασισμένοι στη γραμματική αυτή).

Γραμματικές Χωρίς Συμφραζόμενα

- **Ορισμός. Γραμματική χωρίς συμφραζόμενα** (context-free grammar) είναι μια τετράδα $G = (V, \Sigma, R, S)$ όπου
  - V είναι ένα αλφάβητο
  - Σ, **το σύνολο των τερματικών**, είναι ένα υποσύνολο του V. Τα στοιχεία του $V \setminus \Sigma$ ονομάζονται **μη τερματικά**.
  - R, **το σύνολο των κανόνων (rules or productions)**, είναι ένα πεπερασμένο υποσύνολο του $(V \setminus \Sigma) \times V^*$
  - S, **το αρχικό σύμβολο**, είναι ένα στοιχείο του $V \setminus \Sigma$.
- Ένας κανόνας $(A, u) \in R$ γράφεται συνήθως σαν $A \longrightarrow_G u$ ή $A \longrightarrow u$.

Παράδειγμα

$$G = (V, \Sigma, R, S)$$

- όπου
  - $V = \{S, a, b\}$
  - $\Sigma = \{a, b\}$
  - $R = \{\ S \longrightarrow aSb,\ \ S \longrightarrow e\ \}$
- Οι γραμματικές χωρίς συμφραζόμενα είναι **παραγωγοί γλωσσών.**
- Η παραπάνω γραμματική παράγει τη (μη κανονική) γλώσσα $\{a^n b^n :\ n \geq 0\}$.

Ορισμοί

- Η σχέση $\Longrightarrow_G$ **(παράγει σε ένα βήμα)**:
  Αν $u, v \in V^*$ τότε $u \Longrightarrow_G v$ (ή $u \Longrightarrow v$) ανν υπάρχουν $x, y, v' \in V^*$ και $A \in V \setminus \Sigma$ έτσι ώστε $u = xAy,\ v = xv'y$ και $A \longrightarrow v'$.
- Η σχέση $\Longrightarrow_G^*$ **(παράγει)** είναι η ανακλαστική και μεταβατική κλειστότητα της $\Longrightarrow_G$.
- Κάθε ακολουθία της μορφής

$$w_0 \Longrightarrow_G w_1 \Longrightarrow_G \cdots \Longrightarrow_G w_n$$

  λέγεται **παραγωγή** (derivation) της w_n από την w_0 στην G.
- Μια τέτοια παραγωγή λέμε ότι έχει **μήκος** n (ή αποτελείται από n βήματα).

Ορισμοί

- Αν $G = (V, \Sigma, R, S)$ είναι μια γραμματική χωρίς συμφραζόμενα τότε η **γλώσσα που παράγεται** από την G είναι

$$L(G) = \{w \in \Sigma^* : \ S \Longrightarrow_G^* w\}$$

- Μια γλώσσα L ονομάζεται **γλώσσα χωρίς συμφραζόμενα** (context-free language) ανν $L = L(G)$ όπου G είναι μια γραμματική χωρίς συμφραζόμενα.
- Δυο γραμματικές G_1 και G_2 λέγονται **ισοδύναμες** ανν $L(G_1) = L(G_2)$.

Παράδειγμα

$$G = (V, \Sigma, R, E)$$

- όπου

$$V = \{E, T, F, \mathbf{id}, +, *, (,)\}$$

$$\Sigma = \{\mathbf{id}, +, *, (,)\}$$

και

$$R = \{\ E \longrightarrow T + E, \quad E \longrightarrow T,$$

$$T \longrightarrow F * T, \quad T \longrightarrow F,$$

$$F \longrightarrow (E), \quad F \longrightarrow \mathbf{id}\ \}$$

- Ποιά είναι η γλώσσα που παράγεται από την παραπάνω γραμματική;

Γλώσσες Προγραμματισμού και Γραμματικές

- Όταν χρησιμοποιούμε γραμματικές χωρίς συμφραζόμενα για να ορίσουμε το συντακτικό γλωσσών προγραμματισμού τότε:
  - Τα τερματικά σύμβολα παριστάνουν τα tokens της γλώσσας προγραμματισμού (δηλαδή τις καλά ορισμένες συμβολοσειρές π.χ. if, $x12$, $==$ στην C).
  - Τα μη τερματικά σύμβολα παριστάνουν τα υπόλοιπα δομικά στοιχεία των γλωσσών προγραμματισμού.

Παράδειγμα

$$G = (V, \Sigma, R, S)$$

- όπου

$$V = \{S, (,)\}$$

$$\Sigma = \{(,)\}$$

και

$$R = \{\ S \longrightarrow e, \quad S \longrightarrow SS, \quad S \longrightarrow (S)\ \}$$

- Ποιά είναι η γλώσσα που παράγεται από την παραπάνω γραμματική;

Παράδειγμα

- Δύο παραγωγές της G είναι

$$S \Longrightarrow SS \Longrightarrow S(S) \Longrightarrow S((S)) \Longrightarrow S(()) \Longrightarrow (S)(()) \Longrightarrow ()(())$$

και

$$S \Longrightarrow SS \Longrightarrow (S)S \Longrightarrow ()S \Longrightarrow ()(S) \Longrightarrow ()((S)) \Longrightarrow ()(())$$

- Έτσι βλέπουμε ότι μπορεί η ίδια συμβολοσειρά να έχει περισσότερες από μία παραγωγές σε μία γραμματική χωρίς συμφραζόμενα.

Αριστερότερες Παραγωγές

- Κάθε παραγωγή της μορφής

$$w_0 \Longrightarrow_G w_1 \Longrightarrow_G \cdots \Longrightarrow_G w_n$$

λέγεται **αριστερότερη παραγωγή** (leftmost derivation) ανν το μη τελικό σύμβολο που αντικαθίσταται σε κάθε βήμα είναι το αριστερότερο.

- **Παράδειγμα:** $G = (V, \Sigma, R, S)$ όπου

$$V = \{S, (,)\}$$

$$\Sigma = \{(,)\}$$

και

$$R = \{\ S \longrightarrow e, \quad S \longrightarrow SS, \quad S \longrightarrow (S)\ \}$$

Παράδειγμα

- Η παραγωγή

$$S \Longrightarrow SS \Longrightarrow (S)S \Longrightarrow ()S \Longrightarrow ()(S) \Longrightarrow ()((S)) \Longrightarrow ()(())$$

είναι αριστερότερη, αλλά η

$$S \Longrightarrow S\underline{S} \Longrightarrow S\underline{(S)} \Longrightarrow S((S)) \Longrightarrow S(()) \Longrightarrow (S)(()) \Longrightarrow ()(())$$

δεν είναι.

Αριστερότερες Παραγωγές

- Γράφουμε $\alpha \Longrightarrow_G^L \beta$ όπου $\alpha, \beta \in V^*$ ανν $\alpha = \alpha_1 A \alpha_2,\ \beta = \alpha_1 \gamma \alpha_2,\ A \to \gamma$ είναι ένας κανόνας της G και $\alpha_1 \in \Sigma^*$.
- Η ανακλαστική και μεταβατική κλειστότητα της $\Longrightarrow_G^L$ είναι η $\Longrightarrow_G^{*L}$.

Αριστερότερες Παραγωγές

- **Θεώρημα:** Για κάθε γραμματική χωρίς συμφραζόμενα $G = (V, \Sigma, R, S)$ και κάθε συμβολοσειρά $w \in \Sigma^*$, $S \Longrightarrow_G^* w$ ανν $S \Longrightarrow_G^{*L} w$.

Απόδειξη:

- (Αν) Προφανές.
- (Μόνο αν) Έστω

$$w_0 \Longrightarrow_G w_1 \Longrightarrow_G \cdots \Longrightarrow_G w_n$$

μια παραγωγή όχι υποχρεωτικά αριστερότερη, όπου $w_0 = S$ και $w_n = w \in \Sigma^*$.
Η παραπάνω παραγωγή μπορεί να μετασχηματιστεί σε μια αριστερότερη παραγωγή.

Συνέχεια της Απόδειξης

- Έστω k ο μικρότερος αριθμός τέτοιος ώστε το βήμα $w_k \Longrightarrow_G w_{k+1}$ δεν είναι η αριστερότερη παραγωγή του w_{k+1} από το w_k.
- Τότε η w_k θα είναι της μορφής $\alpha A \beta B \gamma$ όπου $\alpha \in \Sigma^*$, $\beta, \gamma \in V^*$, $A, B \in V \setminus \Sigma$, η w_{k+1} θα είναι της μορφής $\alpha A \beta \delta \gamma$ και υπάρχει ο κανόνας $B \longrightarrow_G \delta$.
- Εφόσον $w_n \in \Sigma^*$, ισχύει $k + 1 < n$ και υπάρχει κάποιο επόμενο βήμα όπου αντικαθίσταται και το A. Έτσι έστω l ο μικρότερος αριθμός, μεγαλύτερος του k έτσι ώστε $w_l = \alpha A \epsilon$ για κάποιο $\epsilon \in V^*$ και $w_{l+1} = \alpha \zeta \epsilon$, όπου $A \longrightarrow_G \zeta$. Τότε $\beta B \gamma \Longrightarrow_G^* \epsilon$ σε $l - k$ βήματα.

Συνέχεια της Απόδειξης

- Αρα μπορούμε να εναλλάξουμε τις χρήσεις των κανόνων $B \longrightarrow_G \delta$ και $A \longrightarrow_G \zeta$ ως εξής.

$$\begin{array}{ll} w_0 \Longrightarrow_G^{*L} w_k = \alpha A \beta B \gamma & k \text{ βήματα} \\ \Longrightarrow_G^{L} \alpha \zeta \beta B \gamma & 1 \text{ βήμα} \\ \Longrightarrow_G^{L} \alpha \zeta \epsilon & l - k \text{ βήματα} \\ \Longrightarrow_G^{L} w_n & n - l - 1 \text{ βήματα} \end{array}$$

- Η νέα παραγωγή έχει μήκος n και αρχίζει με τουλάχιστο $k + 1$ βήματα αριστερής παραγωγής.

Παράδειγμα

- Η παραγωγή

$$S \Longrightarrow \underline{S}S \Longrightarrow S(S) \Longrightarrow S((S)) \Longrightarrow S(()) \Longrightarrow \underline{(S)}(()) \Longrightarrow ()(())$$

μπορεί να γίνει αριστερότερη εφαρμόζοντας το παραπάνω θεώρημα διαδοχικά ως εξής:

- Βήμα 1:

$$S \Longrightarrow SS \Longrightarrow (S)S \Longrightarrow (S)(S) \Longrightarrow (S)((S)) \Longrightarrow (S)(()) \Longrightarrow ()(())$$

- Βήμα 2:

$$S \Longrightarrow SS \Longrightarrow (S)S \Longrightarrow ()S \Longrightarrow ()(S) \Longrightarrow ()((S)) \Longrightarrow ()(())$$

Συντακτικά Δέντρα

- Τα συντακτικά δέντρα (parse trees) είναι μια γραφική μέθοδος αναπαράστασης παραγωγών.
- Θεωρείστε τη γραμματική $G = (V, \Sigma, R, S)$ όπου

$$V = \{S, (,)\}$$

$$\Sigma = \{(,)\}$$

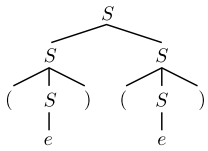
$$R = \{\ S \longrightarrow e,\quad S \longrightarrow SS,\quad S \longrightarrow (S)\ \}$$

Συντακτικά Δέντρα

- Η αριστερότερη παραγωγή

$$S \Longrightarrow SS \Longrightarrow (S)S \Longrightarrow (S)(S) \Longrightarrow ()(S) \Longrightarrow ()()$$

μπορεί να παρασταθεί από το ακόλουθο συντακτικό δέντρο:

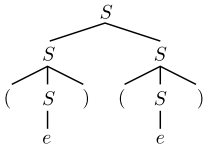


Συντακτικά Δέντρα

- Αλλά και η παραγωγή

$$S \Longrightarrow SS \Longrightarrow S(S) \Longrightarrow (S)(S) \Longrightarrow (S)() \Longrightarrow ()()$$

μπορεί να παρασταθεί από το ίδιο συντακτικό δέντρο:



Συντακτικά Δέντρα

- **Ορισμός.** Ένα δέντρο λέγεται **συντακτικό δέντρο** για την γραμματική χωρίς συμφραζόμενα $G = (V, \Sigma, R, S)$ αν:
  - Οι κόμβοι του δέντρου συμβολίζονται από στοιχεία του συνόλου $V \cup \{e\}$.
  - Η ρίζα του δέντρου συμβολίζεται από το αρχικό σύμβολο S.
  - Κάθε εσωτερικός κόμβος συμβολίζεται με ένα μη τερματικό σύμβολο $A \in V \setminus \Sigma$.
  - Αν ο κόμβος A έχει παιδιά $Q_1, \ldots, X_n$ διαβασμένα από δεξιά πρός αριστερά, τότε

$$A \longrightarrow Q_1 \cdots X_n$$

είναι ένας κανόνας του R.
  - Κάθε κόμβος που συμβολίζεται με e είναι φύλλο και είναι το μοναδικό παιδί του πατέρα του.
- Το **προϊόν (yield)** ενός συντακτικού δέντρου είναι η συμβολοσειρά που προκύπτει με παράθεση από τα αριστερά προς τα δεξιά των συμβόλων της γραμματικής που βρίσκονται στα φύλλα του δέντρου.

Διττές Γραμματικές

- **Ορισμός:** Μια γραμματική G λέγεται **διττή** ανν έχει παραπάνω από μια διαφορετικές αριστερότερες παραγωγές για την ίδια συμβολοσειρά $w \in L(G)$.
- **Παράδειγμα:** Έστω η γραμματική $G = (V, \Sigma, R, E)$ όπου

$$V = \{E, T, F, \mathbf{id}, +, *\}, \quad \Sigma = \{\mathbf{id}, +, *\}$$

και

$$R = \{\ E \longrightarrow E + E, \quad E \longrightarrow E * E, \quad E \longrightarrow \mathbf{id}\ \}$$

Διττές Γραμματικές

- Η γραμματική αυτή έχει δύο διαφορετικές αριστερότερες παραγωγές για τη συμβολοσειρά $\mathbf{id} + \mathbf{id} * \mathbf{id}$:

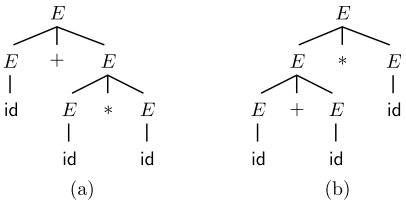
$$E \Longrightarrow E + E \Longrightarrow \mathbf{id} + E \Longrightarrow \mathbf{id} + E * E \Longrightarrow \mathbf{id} + \mathbf{id} * E \Longrightarrow \mathbf{id} + \mathbf{id} * \mathbf{id}$$

και

$$E \Longrightarrow E * E \Longrightarrow E + E * E \Longrightarrow \mathbf{id} + E * E \Longrightarrow \mathbf{id} + \mathbf{id} * E \Longrightarrow \mathbf{id} + \mathbf{id} * \mathbf{id}$$

Διττές Γραμματικές

- Άρα έχουμε και δύο διαφορετικά συντακτικά δέντρα:



Σχέση με τις Κανονικές Γλώσσες

- Το σύνολο των κανονικών γλωσσών είναι υποσύνολο του συνόλου των γλωσσών χωρίς συμφραζόμενα.
- **Παράδειγμα:** Η γραμματική $G = (\{S, a\}, \{a\}, R, S)$ όπου

$$R = \{\ S \longrightarrow aS, \quad S \longrightarrow e\ \}$$

παράγει τη γλώσσα a^*.

Κανονικές Γραμματικές

- **Ορισμός.** Μια γραμματική χωρίς συμφραζόμενα $G = (V, \Sigma, R, S)$ λέγεται **κανονική** ανν

$$R \subseteq (V \setminus \Sigma) \times \Sigma^*((V \setminus \Sigma) \cup \{e\})$$

- **Διαισθητικά:** Μια γραμματική χωρίς συμφραζόμενα λέγεται **κανονική** ανν κάθε κανόνας είναι της μορφής

$$N_1 \longrightarrow wN_2$$

ή

$$N_1 \longrightarrow w$$

όπου w είναι μια συμβολοσειρά αποτελούμενη από τερματικά σύμβολα και N_1, N_2 είναι μη τερματικά σύμβολα.

Παράδειγμα

- Η γραμματική

$$G = (V, \Sigma, R, S)$$

όπου

$$V = \{S, A, B, a, b\}$$

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

$$R = \{\ S \longrightarrow bA, \quad S \longrightarrow aB, \quad S \longrightarrow e,$$

$$A \longrightarrow abaS, \quad B \longrightarrow babS\ \}$$

είναι κανονική.

- Η γλώσσα που παράγεται από την γραμματική αυτή είναι $L(G) = (baba \cup abab)^*$.

Κανονικές Γραμματικές

- **Θεώρημα.** Μια γλώσσα είναι κανονική ανν μπορεί να παραχθεί από μια κανονική γραμματική.

Απόδειξη: (Μόνο αν)

- Έστω μια κανονική γλώσσα που γίνεται δεκτή από το αυτόματο $M = (K, \Sigma, \delta, s, F)$. Τότε $L(M) = L(G)$ όπου $G = (V, \Sigma, R, S)$ είναι η κανονική γραμματική που ορίζεται ως εξής:
  - $V = \Sigma \cup K$
  - $S = s$
  - $R = \{q \longrightarrow ap : \ \delta(q, a) = p\} \cup \{q \longrightarrow e : \ q \in F\}$

Συνέχεια της Απόδειξης

- Οι κανόνες της G έχουν σχεδιαστεί ακριβώς για να μιμούνται τις κινήσεις του M.
- Δηλαδή για κάθε $\sigma_1, \ldots, \sigma_n \in \Sigma$ και $p_0, \ldots, p_n \in K$,

$$(p_0, \sigma_1 \cdots \sigma_n) \vdash_M (p_1, \sigma_2 \cdots \sigma_n) \vdash_M \cdots \vdash_M (p_n, e)$$

ανν

$$p_0 \Longrightarrow_G \sigma_1 p_1 \Longrightarrow_G \sigma_1 \sigma_2 p_2 \cdots \Longrightarrow_G \sigma_1 \cdots \sigma_n p_n$$

Συνέχεια της Απόδειξης

- Ας υποθέσουμε ότι $w \in L(M)$. Τότε $(s, w) \vdash_M^* (p, e)$ για κάποιο $p \in F$. Τότε $s \Longrightarrow^* wp$ και εφόσον $p \to e$ είναι επίσης κανόνας της G, θα ισχύει $s \Longrightarrow^* w$ και $w \in L(G)$. Αρα $L(M) \subseteq L(G)$.
- Ας υποθέσουμε ότι $w \in L(G)$. Τότε $s \Longrightarrow_G^* w$. Ο κανόνας που χρησιμοποιήθηκε στο τελευταίο βήμα της παραγωγής πρέπει να ήταν της μορφής $p \longrightarrow e$, για κάποιο $p \in F$, έτσι ώστε $s \Longrightarrow_G^* wp \Longrightarrow_G^* w$. Όμως τότε $(s, w) \vdash_M^* (p, e)$ και $w \in L(M)$. Αρα $L(G) \subseteq L(M)$.
- Τελικά $L(G) = L(M)$.

Συνέχεια της Απόδειξης

(Αν)

- Έστω $G = (V, \Sigma, R, E)$ μία κανονική γραμματική. Ορίζουμε ένα μη ντετερμινιστικό πεπερασμένο αυτόματο M έτσι ώστε $L(M) = L(G)$.
- Έστω $M = (K, \Sigma, \Delta, s, F)$, όπου

$$K = (V \setminus \Sigma) \cup \{f\} \quad (f \text{ είναι ένα νέο σύμβολο})$$

$$s = S$$

$$F = \{f\}$$

$$\Delta = \{(A, w, B) : \ A \longrightarrow wB \in R \text{ και } A, B \in V \setminus \Sigma \text{ και } w \in \Sigma^*\}$$

$$\cup \{(A, w, f) : \ A \to w \in R \text{ και } A \in V \setminus \Sigma \text{ και } w \in \Sigma^*\}.$$

Συνέχεια της Απόδειξης

- Οι κινήσεις του M μιμούνται τις παραγωγές της G. Για κάθε $A_1, \ldots, A_n \in V \setminus \Sigma, w_1, \ldots, w_n \in \Sigma^*$,

$$A_1 \Longrightarrow_G w_1 A_2 \Longrightarrow_G w_1 w_2 A_3 \cdots$$

$$\cdots \Longrightarrow_G w_1 w_2 \ldots w_{n-1} A_n \Longrightarrow_G w_1 w_2 ... w_{n-1} w_n$$

ανν

$$(A_1, w_1 \cdots w_n) \vdash_M (A_2, w_2 \cdots w_n) \vdash_M \cdots \vdash_M (A_n, w_n) \vdash_M (f, e).$$

Συνέχεια της Απόδειξης

- Επομένως, αν $w \in L(G)$ τέτοια ώστε $w \in \Sigma^*$ και $S \Longrightarrow_G^* w$, τότε $(S, w) \vdash_M^* (f, e)$ και έτσι $w \in L(M)$. Αρα $L(G) \subseteq L(M)$.
- Αντίστροφα , αν $w \in L(M)$ τέτοια ώστε $(S, w) \vdash_M^* (f, e)$ τότε $S \Longrightarrow_G^* w$, έτσι ώστε $w \in L(G)$. Αρα $L(M) \subseteq L(G)$.
- Τελικά $L(M) = L(G)$.

Μελέτη

- Κεφάλαια 3.1-3.2 από το βιβλίο:
  Harry R. Lewis και Χρίστος Χ. Παπαδημητρίου, Στοιχεία Θεωρίας Υπολογισμού, Τεχνικό Επιμελητήριο Ελλάδας, 1992.