

Κανονικές Γλώσσες

Αφού ορίσουμε τις έννοιες του αλφαβήτου και της γλώσσας, θα δούμε τι πράξεις μπορούμε να κάνουμε μεταξύ γλωσσών και θα διαπιστώσουμε πως συχνά μπορούμε να περιγράψουμε με συντομία μία γλώσσα ακόμα κι αν αυτή είναι άπειρη.

Συγκεκριμένα, θα μελετήσουμε μια ειδική κατηγορία γλωσσών, τις κανονικές γλώσσες, και θα τις συσχετίσουμε με τις κανονικές εκφράσεις. Αργότερα, θα δούμε πως οι κανονικές γλώσσες είναι ακριβώς οι γλώσσες που γίνονται αποδεκτές από τα απλά υπολογιστικά μοντέλα που ονομάζουμε πεπερασμένα αυτόματα.

Αλφάβητο – Γλώσσα – Πράξεις Γλωσσών

Αλφάβητο: ένα οποιοδήποτε πεπερασμένο σύνολο συμβόλων, παραδείγματα αλφαβήτων είναι τα $\{0,1\}$, $\{\alpha, \beta, \gamma, \delta\}$, και $\{\&, \pi, *, \mathbb{E}\}$.

Συνήθως χρησιμοποιούμε το γράμμα Σ για να συμβολίσουμε ένα αλφάβητο, ενώ τα αλφάβητα που κυρίως θα χρησιμοποιούμε θα είναι το $\{0,1\}$, και το $\{a,b\}$.

Συμβολοσειρά ή λέξη του Σ : μια πεπερασμένη ακολουθία συμβόλων του Σ . Π.χ. αν $\Sigma = \{0,1\}$, τα 0, 0010, 101010101 είναι συμβολοσειρές του Σ .

Μήκος μιας συμβολοσειράς x είναι το πλήθος των συμβόλων της, και το συμβολίζουμε με $|x|$. Π.χ. αν $x = abbbab$, τότε $|x| = 6$.

Τη συμβολοσειρά που δεν έχει κανένα σύμβολο, την ονομάζουμε κενή συμβολοσειρά, και τη συμβολίζουμε με ε . Προφανώς, $|\varepsilon| = 0$.

Μπορούμε να ενώσουμε δύο συμβολοσειρές x, y με την πράξη της συνένωσης (ή παράθεσης) την οποία συμβολίζουμε με $x \circ y$ ή απλά xy . Π.χ. αν $x = abb$, $y = abaa$ και $z = \varepsilon$, τότε $xy = abbabaa$, $yx = abaaabb$, και $zx = xz = abb$.

Το σύνολο όλων των συμβολοσειρών του Σ το συμβολίζουμε Σ^* . Π.χ. αν $\Sigma = \{0,1\}$, τότε $\Sigma^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, \dots\}$. Το σύμβολο $*$ το ονομάζουμε αστέρι Kleene¹ και θα το συναντήσουμε και στη συνέχεια. Συμβολίζει την επιλογή οσονδήποτε συμβόλων του συνόλου στο οποίο εφαρμόζεται (μηδέν ή περισσότερα σύμβολα).

Γλώσσα πάνω στο Σ : ένα οποιοδήποτε υποσύνολο του Σ^* . Μια γλώσσα L μπορεί να είναι πεπερασμένη ή άπειρη. Παραδείγματα γλωσσών για $\Sigma = \{0,1\}$:

$$L_1 = \{x \mid \text{η } x \text{ έχει περιττό αριθμό } 0\}$$

$$L_2 = \emptyset$$

$$L_3 = \{\varepsilon\}$$

$$L_4 = \{x \mid \text{η } x \text{ αρχίζει με } 00 \text{ και περιέχει το } 1001\}$$

$$L_5 = \{x \mid \text{η } x \text{ είναι παλινδρομική συμβολοσειρά μήκους το πολύ } 99\}$$

Εφόσον οι γλώσσες είναι σύνολα, ισχύουν για αυτές οι γνωστές συνολοθεωρητικές πράξεις (τομή, ένωση, συμπλήρωμα, διαφορά). Παρακάτω, θυμίζουμε τις πράξεις αυτές και εισάγουμε και κάποιες καινούργιες, όπως η συνένωση και το αστέρι Kleene.

Έστω L_1, L_2, L γλώσσες πάνω σε ένα αλφάβητο Σ .

Τομή: $L_1 \cap L_2 = \{x \mid x \in L_1 \text{ και } x \in L_2\}$

Ένωση: $L_1 \cup L_2 = \{x \mid x \in L_1 \text{ ή } x \in L_2\}$

Συμπλήρωμα: $L' = \{x \mid x \notin L\}$

Συνένωση: $L_1 L_2 = \{xy \mid x \in L_1 \text{ και } y \in L_2\}$,

Με βάση τη συνένωση ορίζουμε και την $L^k = \underbrace{LL \dots L}_{k \text{ όροι}}$

Αστέρι Kleene: $L^* = \{x \mid \eta \text{ } x \text{ είναι συνένωση } 0 \text{ ή περισσότερων συμβολοσειρών της } L\}$

Σχετικός είναι και ο παρακάτω συμβολισμός που χρησιμοποιούμε ενίοτε:

$L^+ = \{x \mid \eta \text{ } x \text{ είναι συνένωση μίας ή περισσότερων συμβολοσειρών της } L\}$

Παράδειγμα 1. Έστω οι γλώσσες πάνω στο $\{0, 1\}$: $L_1 = \{x \mid \eta \text{ } x \text{ έχει περιττό αριθμό } 0\}$ και $L_2 = \{x \mid \eta \text{ } x \text{ αρχίζει από } 1\}$ και οι συμβολοσειρές $x = 01010, y = 110, z = 0000$.

Να διαπιστώσετε αν τα x, y, z ανήκουν ή όχι στις γλώσσες, $L_1, L_2, L_1 \cap L_2, L_1 \cup L_2, L_1 L_2, L_2 L_1$ και L_1^* .

Η x έχει περιττό αριθμό 0 και δεν αρχίζει από 1, άρα $x \in L_1, x \notin L_2$. Επομένως, $x \in L_1 \cup L_2$, και $x \notin L_1 \cap L_2$. Παρατηρούμε επίσης ότι η x γράφεται ως $0 \circ 1010$ και $0 \in L_1, 1010 \in L_2$. Άρα $x \in L_1 L_2$. Από την άλλη, η x δεν γίνεται να γραφτεί ως yw με την y να αρχίζει από 1 και άρα $x \notin L_2 L_1$. Τέλος αφού το $x \in L_1$, θα ισχύει επίσης ότι $x \in L_1^*$.

Η y έχει περιττό αριθμό 0 και αρχίζει από 1, άρα $y \in L_1, y \in L_2$. Επομένως, $y \in L_1 \cup L_2, y \in L_1^*$, και $y \in L_1 \cap L_2$. Επίσης, $y = 1 \circ 10$, και άρα $y \in L_2 L_1$. Από την άλλη, η y δεν γίνεται να γραφτεί ως xw με την x να αρχίζει έχει περιττό αριθμό 0 και την w να αρχίζει από 1, και άρα $x \notin L_1 L_2$.

Η z έχει άρτιο αριθμό 0 και δεν αρχίζει από 1, άρα $z \notin L_1, z \notin L_2$. Επομένως, $z \notin L_1 \cup L_2, z \notin L_1 \cap L_2$. Η z δεν περιέχει καθόλου 1 άρα δεν μπορεί να είναι αποτέλεσμα συνένωσης με λέξη της L_2 . Επομένως $z \notin L_1 L_2$, και $z \notin L_2 L_1$. Παρατηρούμε όμως ότι η z γράφεται ως $0 \circ 000$ και $0 \in L_1, 000 \in L_1$. Άρα $z \in L_1^*$. $\square$

Κανονικές Εκφράσεις & Κανονικές Γλώσσες

Θα θέλαμε έναν εύκολο και συνοπτικό τρόπο να περιγράψουμε τον τρόπο που προκύπτουν οι λέξεις μιας γλώσσας. Αυτό δεν είναι πάντα δυνατό. Στις γλώσσες που θα μελετήσουμε εδώ, πάντως, μπορούμε να περιγράψουμε κάθε γλώσσα με μία κανονική έκφραση. Πριν δώσουμε τους ακριβείς ορισμούς ας αρχίσουμε με ένα παράδειγμα:

Έστω η γλώσσα $L = \{x \in \{a, b\}^* \mid x \text{ αρχίζει με } a \text{ και περιέχει το } bbb\}$. Την L την περιγράφουμε με την κανονική έκφραση $a(a+b)^*bbb(a+b)^*$. Το "+" σημαίνει "ή" και το "*" σημαίνει παράθεση κανενός ή περισσότερων συμβόλων. Επομένως η έκφραση $a(a+b)^*bbb(a+b)^*$ στα ελληνικά γίνεται: «η συμβολοσειρά αρχίζει με a , ακολουθούν οσαδήποτε a ή b , ακολουθεί το bbb και τέλος ακολουθούν οσαδήποτε a ή b ».

Δίνουμε τώρα τον αναδρομικό ορισμό του τι είναι κανονική έκφραση καθώς και τον ορισμό της κανονικής γλώσσας:

Μια κανονική έκφραση πάνω σε ένα αλφάβητο Σ ορίζεται αναδρομικά ως εξής:

1. $\emptyset$ είναι η κανονική έκφραση που αντιστοιχεί στην κενή γλώσσα $\emptyset$.
2. ε είναι η κανονική έκφραση που αντιστοιχεί στη γλώσσα $\{\varepsilon\}$.
3. Για κάθε σύμβολο $a \in \Sigma$, a είναι η κανονική έκφραση που αντιστοιχεί στη γλώσσα $\{a\}$.
4. Αν r και s είναι εκφράσεις που αντιστοιχούν στις γλώσσες L_r και L_s , τότε και οι (rs) , $(r+s)$ και (r^*) είναι κανονικές εκφράσεις που αντιστοιχούν στις γλώσσες $L_r L_s$, $L_r \cup L_s$ και L_r^* .
5. Τίποτα άλλο δεν είναι κανονική έκφραση, εκτός αν προκύπτει από τους παραπάνω κανόνες.

Κανονική γλώσσα θα ονομάζεται κάθε γλώσσα για την οποία υπάρχει κανονική έκφραση που την περιγράφει.

Συνεπώς για να δείξουμε ότι μια γλώσσα είναι κανονική αρκεί να βρούμε μια κανονική έκφραση για τη γλώσσα!

Παρατηρήσεις:

α) Οι πράξεις + (ένωση), * (αστέρι Kleene) και $\circ$ (συνένωση, της οποίας το σύμβολο κατά κανόνα παραλείπουμε) έχουν κάποια προτεραιότητα η οποία μας επιτρέπει να παραλείπουμε κάποιες από τις παρενθέσεις. Προηγείται το *, ακολουθεί η συνένωση, και τέλος η +. Επιπλέον, η ένωση και η συνένωση είναι προσεταιριστικές. Επομένως γράφουμε την έκφραση $0 + 1^*0(1 + 0)$, και εννοούμε $\left(0 + \left(\left(1^*0\right)(1 + 0)\right)\right)$.

β) Οι κανονικές γλώσσες είναι κατά κάποιο τρόπο οι απλούστερες γλώσσες που μπορούμε να κατασκευάσουμε ή να αναγνωρίσουμε. Θα δούμε στη συνέχεια πως υπάρχουν και μη κανονικές γλώσσες.

γ) Μια κανονική έκφραση περιγράφει ακριβώς μία κανονική γλώσσα, αλλά μία κανονική γλώσσα μπορεί να περιγραφεί με άπειρες διαφορετικές κανονικές εκφράσεις. Π.χ. η έκφραση $0(0+1)^*$ περιγράφει τη γλώσσα με όλες τις συμβολοσειρές που αρχίζουν από 0. Την ίδια γλώσσα όμως περιγράφει και η $0(\varepsilon+1)0^*(0+1)^*$.

Παράδειγμα 2. Να δώσετε κανονικές εκφράσεις για τις ακόλουθες κανονικές γλώσσες:

α) $L_1 = \{x \in \{0,1\}^* \mid \text{η } x \text{ αρχίζει με } 0 \text{ και το προτελευταίο της σύμβολο είναι } 1\}$

β) $L_2 = \{x \in \{0,1\}^* \mid \text{η } x \text{ έχει μήκος το πολύ } 34\}$

γ) $L_3 = \{x \in \{0,1\}^* \mid \text{η } x \text{ έχει μήκος τουλάχιστον } 3\}$

δ) $L_4 = \{x \in \{a,b\}^* \mid \text{η } x \text{ περιέχει το } aa\}$

ε) $L_5 = \{x \in \{a,b\}^* \mid \text{η } x \text{ δεν περιέχει το } aa\}$

στ) $L_6 = \{x \in \{0,1\}^* \mid \text{η } x \text{ αναπαριστά δυαδικό αριθμό που δεν διαιρείται με το } 2\}$

ζ) $L_7 = \{x \in \{0,1\}^* \mid \text{η } x \text{ δεν τελειώνει σε } 00\}$

η) $L_8 = \{x \in \{0,1\}^* \mid \text{η } x \text{ έχει μήκος πολλαπλάσιο του } 3 \text{ ή τελειώνει σε } 1111\}$

θ) $L_9 = \{x \in \{0,1\}^* \mid \text{η } x \text{ έχει άρτιο πλήθος } 1\}$

α) $0(0+1)^*1(0+1)$

β) Αν θέλαμε μήκος ακριβώς 34, η κ.ε. θα ήταν $(0+1)^{34}$. Τώρα θέλουμε τη δυνατότητα κάποιες από τις 34 παρενθέσεις να μην συνεισφέρουν σύμβολο. Επομένως η ζητούμενη κ.ε. είναι η $(\varepsilon+0+1)^{34}$.

γ) $(0+1)(0+1)(0+1)(0+1)^*$

δ) $(a+b)^*aa(a+b)^*$

ε) Το γεγονός ότι η λέξη δεν περιέχει aa σημαίνει ότι αν εμφανίζονται μονά a αυτά θα είναι είτε στην αρχή της λέξης, είτε στο τέλος, είτε στη μέση με τουλάχιστον ένα b πριν και ένα μετά. Μπορούμε να φτιάξουμε πολλές διαφορετικές κ.ε., όπως η $b^* + (b^*ab)^*b^* + b^*(bab^*)^*$, ή η $(a+\varepsilon)(b^+ab^+)^*b^*(a+\varepsilon)$.

στ) $1 + 1(0+1)^*1$

ζ) $(0+1)^*(01+10+11)+0+1+\varepsilon$

η) $((0+1)(0+1)(0+1))^* + (0+1)^*1111$

θ) $(0^*10^*10^*)^*$ ή $(0^*10^*1)^*0^*$

□

Παράδειγμα 3. Να αναγνωρίσετε τις κανονικές γλώσσες. στις οποίες αντιστοιχούν οι παρακάτω κανονικές εκφράσεις:

α) $(a+b)^*a(a+b)^*$

β) b^*ab^*

γ) $(0+1)^*110(0+1)^*$

δ) $((a+b)(a+b))^*a$

ε) $(a+\varepsilon)b^* + aaa$

στ) $(00+01+11)(0+1)^*0$

$$\zeta) (0 + 1)^5(0 + 1 + \varepsilon)^8$$

$$\eta) (1^*01^*01^*0)^*$$

$$\alpha) L_1 = \{x \in \{a, b\}^* \mid \text{η } x \text{ περιέχει τουλάχιστον ένα } a\}$$

$$\beta) L_2 = \{x \in \{a, b\}^* \mid \text{η } x \text{ περιέχει ακριβώς ένα } a\}$$

$$\gamma) L_3 = \{x \in \{0, 1\}^* \mid \text{η } x \text{ περιέχει το } 110\}$$

$$\delta) L_4 = \{x \in \{a, b\}^* \mid \text{η } x \text{ έχει περιττό μήκος και τελειώνει σε } a\}$$

$$\varepsilon) L_5 = \{x \in \{a, b\}^* \mid \text{η } x \text{ είναι η } aaa, \text{ ή έχει μόνο } b \text{ ή έχει } 1 \text{ } a \text{ στην αρχή και μετά } b\}$$

$$\sigma\tau) L_6 = \{x \in \{0, 1\}^* \mid \text{η } x \text{ έχει μήκος } \geq 3, \text{ δεν αρχίζει με } 10 \text{ και τελειώνει σε } 0\}$$

$$\zeta) L_7 = \{x \in \{0, 1\}^* \mid \text{η } x \text{ έχει μήκος τουλάχιστον } 5 \text{ και το πολύ } 13\}$$

$$\eta) L_8 = \{x \in \{0, 1\}^* \mid \text{η } x \text{ λήγει σε } 0 \text{ και το πλήθος των } 0 \text{ είναι πολ/σιο του } 3\} \quad \square$$

Θα κλείσουμε με δύο σημαντικά θεωρήματα για τις κανονικές γλώσσες.

Θεώρημα 1. Κάθε πεπερασμένη γλώσσα είναι κανονική.

Απόδειξη: Έστω $L = \{x_1, x_2, \dots, x_k\}$ μία πεπερασμένη γλώσσα. Εφόσον η L περιγράφεται από την κανονική έκφραση $x_1 + x_2 + \dots + x_k$, είναι κανονική. (Αν θέλαμε να είμαστε πιο αυστηροί θα αποδεικνύαμε το θεώρημα με επαγωγή.) $\square$

Πριν διατυπώσουμε το επόμενο θεώρημα, θα εισάγουμε μία ακόμη έννοια. Θα λέμε ότι οι κανονικές γλώσσες είναι *κλειστές*² ως προς την ένωση (αντίστοιχα τομή, συνένωση κτλ.), αν κάθε φορά που L_1, L_2 είναι κανονικές, το ίδιο ισχύει για την $L_1 \cup L_2$ (αντίστοιχα $L_1 \cap L_2, L_1 L_2$ κτλ.).

Θεώρημα 2. Οι κανονικές γλώσσες είναι κλειστές ως προς την ένωση, τη συνένωση και το αστέρι Kleene.

Στην πραγματικότητα, οι κανονικές γλώσσες είναι κλειστές και ως προς την τομή, και ως προς το συμπλήρωμα, αλλά αυτό θα το δούμε αργότερα.

Ασκήσεις

Άσκηση 1. Να δώσετε κανονικές εκφράσεις για τις ακόλουθες κανονικές γλώσσες:

$$\alpha) L_1 = \{x \in \{0, 1\}^* \mid \text{η } x \text{ αρχίζει με } 0 \text{ και περιέχει το } 001\}$$

$$\beta) L_2 = \{x \in \{0, 1\}^* \mid \text{η } x \text{ έχει μήκος τουλάχιστον } 2 \text{ και το πολύ } 8\}$$

$$\gamma) L_3 = \{x \in \{a, b\}^* \mid \text{η } x \text{ δεν περιέχει το } aa \text{ ή τελειώνει σε } bb\}$$

$$\delta) L_4 = \{x \in \{a, b\}^* \mid \text{η } x \text{ περιέχει και το } aa \text{ και το } bb\}$$

$$\varepsilon) L_5 = \{x \in \{0, 1\}^* \mid \text{η } x \text{ δεν αρχίζει με } 10\}$$

² Πιο σωστά, θα λέμε ότι το σύνολο των κανονικών γλωσσών είναι κλειστό...

στ) $L_6 = \{x \in \{0,1\}^* \mid \eta \ x \text{ αναπαριστά δυαδικό αριθμό που διαιρείται με το } 2\}$

ζ) $L_7 = \{x \in \{a,b\}^* \mid \eta \ x \text{ περιέχει ακριβώς } 2 \ a\}$

η) $L_8 = \{x \in \{0,1\}^* \mid \eta \ x \text{ έχει μήκος πολλαπλάσιο του } 2\}$

θ) $L_9 = \{x \in \{0,1\}^* \mid \eta \ x \text{ περιέχει περιττό αριθμό } 1 \text{ ή άρτιο αριθμό } 0\}$

ι) $L_{10} = \{x \in \{a,b\}^* \mid \eta \ x \text{ περιέχει τουλάχιστον } 2 \ a\}$

Άσκηση 2. Να αναγνωρίσετε τις κανονικές γλώσσες. στις οποίες αντιστοιχούν οι παρακάτω κανονικές εκφράσεις:

α) $(a + b)^* ab(a + b)^* ab(a + b)^*$

β) $(b^* b(a + aa)bb^*)^*$

γ) $10(0 + 1)^* 01$

δ) $(1^* 01^* 0)^* 1^*$

ε) $(1^* 01^* 0)^*$

στ) $bbb((a + b)(a + b)(a + b))^*$

ζ) $(0 + 1 + \varepsilon)^{99}(0 + 1)$

Άσκηση 3. Να αποδείξετε το Θεώρημα 2.