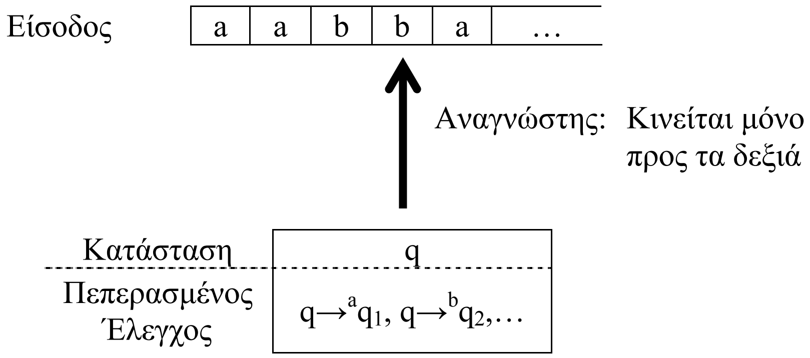
Κεφάλαιο 2: Γλώσσες Τύπου 3 (Κανονικές Γλώσσες)

2.1 Πεπερασμένα Αυτόματα

Λίγα γενικά λόγια:

- Μηχανισμός αναγνώρισης γλώσσας
- Χωρίς βοηθητική μνήμη
- Απλός αλλά χρήσιμος σε αρκετές εφαρμογές, πχ στην λεκτική ανάλυση, μέρος των compilers



Ντετερμινιστικό Πεπερασμένο Αυτόματο

Πεντάδα $A=(Q, \Sigma, \delta, q_0, F)$

- Q: πεπερασμένο σύνολο από *καταστάσεις*
- Σ: πεπερασμένο *αλφάβητο*
- $q_0 \in Q$: *αρχική κατάσταση*
- $F \subseteq Q$: σύνολο των *τελικών καταστάσεων*
- δ: $Q\times\Sigma\rightarrow Q$: *συνάρτηση μετάβασης*

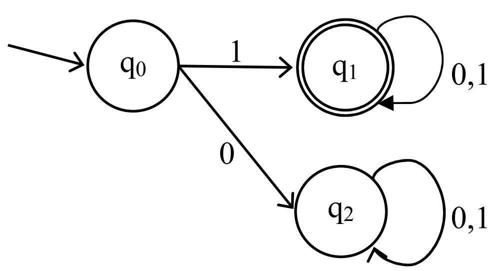
➢ Η συνάρτηση μετάβασης μπορεί να επεκταθεί σε μία συνάρτηση με είσοδο λέξεις με προφανή τρόπο:
δ^*: $Q\times\Sigma^*\rightarrow Q$, όπου:
  - ✓ $\delta^*(q,\varepsilon)=q$
  - ✓ $\delta^*(q,wx)=\delta(\delta^*(q,w),x)$, για $x\in\Sigma$, $w\in\Sigma^*$

➢ $L(A)=\{w\in\Sigma^* \mid \delta^*(q_0,w)\in F\}$
Η L(A) είναι η γλώσσα *που γίνεται δεκτή* από το A

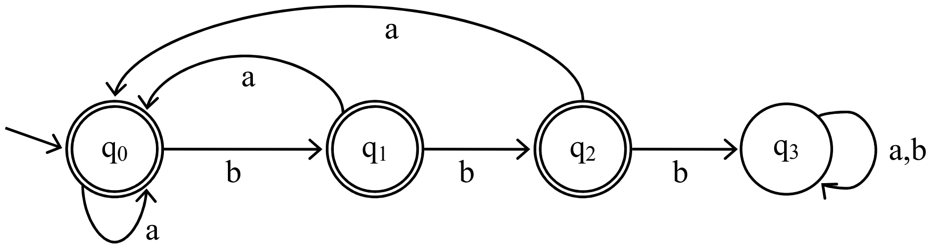
Έστω $w=a_1a_2\ldots a_n\in\Sigma^*$ και $q_0q_1\ldots q_n\in Q^*$ με $q_{i+1}=\delta(q_i,a_{i+1})$ $i=0,1,\ldots,n-1$.
$q_0q_1\ldots q_n$ είναι η *διαδρομή* που *αντιστοιχεί* στο $a_1a_2\ldots a_n$.
Εάν $q_n\in F$ η διαδρομή είναι *αποδεχόμενη*, αλλιώς *μη αποδεχόμενη*.

<u>Παραδείγματα</u>:

1. $L(A)=\{1u \mid u\in\{0,1\}^*\}=1\{0,1\}^*$



2. Πεπερασμένο αυτόματο A με $L(A)=\{w\in\{a,b\}^* \mid w$ δεν περιέχει 3 συνεχόμενα $b\}$



Μη Ντετερμινιστικό Πεπερασμένο Αυτόματο

$A=(Q, \Sigma, \delta, q_0, F)$

- δ: πεπερασμένο υποσύνολο του $Q\times\Sigma\times Q$ (*σχέση μετάβασης*)
  Ισοδύναμα: $\delta: Q\times\Sigma\rightarrow 2^Q$
- Και πάλι επέκταση:
  - ✓ $\delta^*(q,\varepsilon)=\{q\}\ \forall q\in Q$
  - ✓ $\delta^*(q,wx)=\cup_{q'\in\delta^*(q,w)}\delta(q',x)$, για $x\in\Sigma$, $w\in\Sigma^*$

- $L(A)=\{w\in\Sigma^* \mid \delta^*(q_0,w)\cap F\neq\emptyset\}$

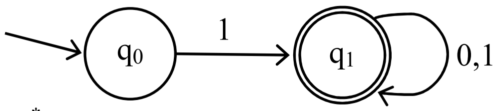
Διαδρομές

Έστω $w=a_1a_2\ldots a_n\in\Sigma^*$ και $q_0q_1\ldots q_m\in Q^*$ με $(q_i, a_{i+1}, q_{i+1})\in\delta$.

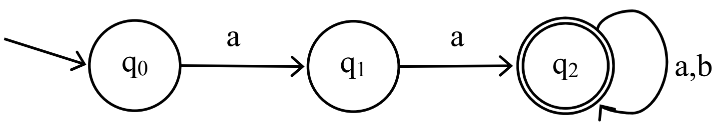
- Αν m=n τότε η $q_0q_1\ldots q_n$ λέγεται *διαδρομή που αντιστοιχεί* στο $a_1a_2\ldots a_n$.
  Αν $q_n\in F$ τότε η διαδρομή λέγεται *αποδεχόμενη*, αλλιώς *μη αποδεχόμενη*.
- Αν m<n και $\delta(q_m, a_{m+1})=\emptyset$ τότε η $q_0q_1\ldots q_m\perp$ λέγεται *μη αποδεχόμενη διαδρομή*.

Παραδείγματα:

1. $L(A)=\{1u \mid u\in\{0,1\}^*\}$



2. $L(A)=\{aau \mid u\in\{a,b\}^*\}$



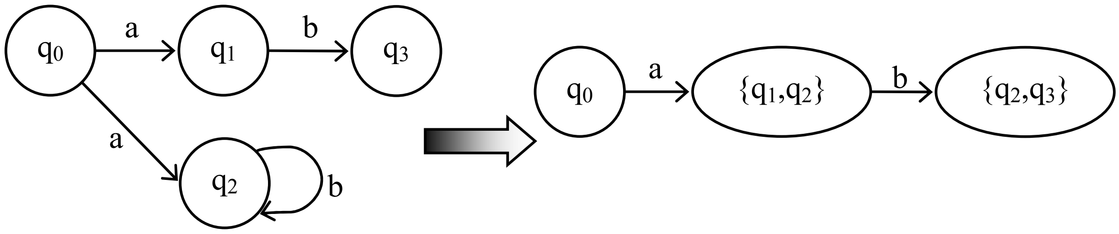
2.2 Ισοδυναμία Ντετερμινιστικών και μη Ντετερμινιστικών Πεπερασμένων Αυτομάτων

Προφανώς τα ντετερμινιστικά πεπερασμένα αυτόματα (*ΝΠΑ*) είναι ειδική περίπτωση των μη ντετερμινιστικών πεπερασμένων αυτομάτων (*ΜΠΑ*). Θα δείξουμε όμως ότι τα ΜΠΑ δεν μπορούν να αποδεχτούν καμία γλώσσα που δεν την αποδέχεται ένα ΝΠΑ.

Θεώρημα 2.1 Για κάθε ΜΠΑ A_1 υπάρχει ένα ΝΠΑ A_2 έτσι ώστε $L(A_1)=L(A_2)$. Τότε, τα A_1, A_2 ονομάζονται *ισοδύναμα*.

Απόδειξη

Έστω ένα ΜΠΑ $A_1=(Q, \Sigma, \delta, q_0, F)$. Θα κατασκευάσουμε ένα ισοδύναμο ΝΠΑ A_2. Η κεντρική ιδέα της κατασκευής είναι να θεωρήσουμε ότι ένα ΜΠΑ δεν βρίσκεται σε μία κατάσταση, αλλά σε ένα <u>σύνολο καταστάσεων</u>. Συγκεκριμένα, σε όλες τις καταστάσεις που μπορεί να οδηγηθεί από την αρχική κατάσταση, «καταναλώνοντας» την είσοδο που έχει διαβάσει ως τώρα:



Τώρα θα δείξουμε την μαθηματική τυποποίηση αυτής της ιδέας.
Θέτουμε: $A_2=\{2^Q, \Sigma, \Delta, \{q_0\}, F'\}$, όπου:

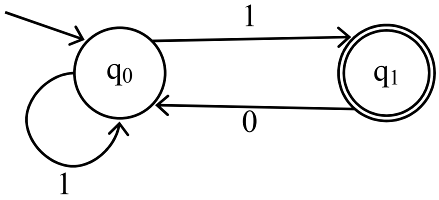
- 2^Q είναι τα υποσύνολα του Q, δηλαδή τα υποσύνολα των καταστάσεων του A_1
- $F'=\{P\subseteq Q \mid P\cap F\neq\varnothing\}$
- $\Delta(P, \alpha)=\cup_{p\in P}\delta(p,\alpha)$

Ισχύει: $\Delta^*(\{q\},w)=\delta^*(q,w)$ (*)

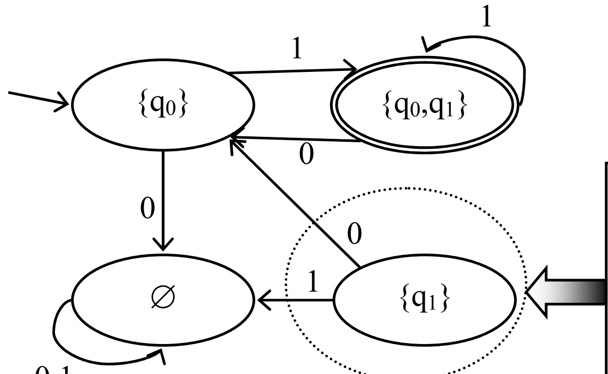
Απόδειξη της σχέσης (*)
Με επαγωγή ως προς w.
Αν w=ε: $\delta^*(q,w)=\{q\}=\Delta^*(\{q\},w)$.
Έστω ότι η (*) ισχύει για το $w\in\Sigma^*$ και έστω $\alpha\in\Sigma$.
Τότε: $\delta^*(q,w\alpha)=$
$=\cup_{t\in\delta^*(q,w)}\delta(t,\alpha)=$
$=\cup_{t\in\Delta^*(\{q\},w)}\delta(t,\alpha)=$
$=\Delta(\Delta^*(\{q\},w),\alpha)=$
$=\Delta^*(\{q\},w\alpha)$.

Οπότε: $w\in L(A_1)\Leftrightarrow$
$\Leftrightarrow\delta^*(q_0,w)\cap F\neq\varnothing\Leftrightarrow$
$\Leftrightarrow\Delta^*(\{q_0\},w)\in F'\Leftrightarrow$
$\Leftrightarrow w\in L(A_2)$.

<u>Παραδείγματα</u>:
Το ΜΠΑ A_1:

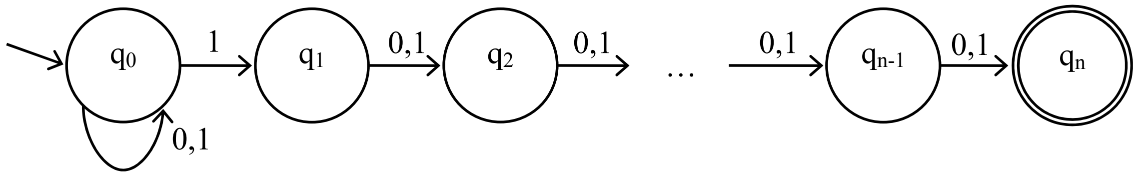


Δίνει το ΝΠΑ A_2:



<u>Παρατήρηση</u>:
Άχρηστο μέρος διότι ποτέ δεν μπορούμε να φτάσουμε στο $\{q_1\}$ από το $\{q_0\}$.
Η παρουσία του δεν αλλάζει τίποτα.

Παρατήρηση:
Το A_2 έχει εκθετικό μέγεθος ως προς το A_1.
Δεν είναι πρόβλημα μόνο της συγκεκριμένης κατασκευής, αλλά γενικότερο: υπάρχουν γλώσσες για τις οποίες τα ΜΠΑ είναι πολύ μικρότερα από τα ΝΠΑ.
Για παράδειγμα:
$L_n=\{w\in\{0,1\}^* \mid$ το n-στο σύμβολο πριν το τέλος είναι 1$\}$.
Αν το ΜΠΑ είναι:



Τότε κάθε ΝΠΑ A με $L(A)=L_n$ έχει εκθετικό αριθμό καταστάσεων (χωρίς απόδειξη).

2.3 Πεπερασμένα Αυτόματα και Γλώσσες Τύπου 3

Λήμμα 2.1 Για κάθε ΝΠΑ A υπάρχει μία γραμματική G τύπου 3 τέτοια ώστε L(G)=L(A).
Απόδειξη
Έστω $A=(Q, \Sigma, \delta, q_0, F)$.
Ορίζουμε $G=(Q, \Sigma, P, q_0)$:
- Αν $q_0\in F \Rightarrow (q_0 ::= \varepsilon)\in P$
- Αν $\delta(q,\alpha)=p \Rightarrow (q ::= \alpha p)\in P$
- Αν $\delta(q,\alpha)=p\in F \Rightarrow (q ::= \alpha)\in P$

Μία εύκολη επαγωγή δείχνει ότι L(G)=L(A).

Λήμμα 2.2 Για κάθε γραμματική G τύπου 3 υπάρχει ένα ΜΠΑ A με L(G)=L(A).
(Χωρίς απόδειξη)

Θεώρημα 2.2 Για κάθε γλώσσα L ισχύει ότι L=L(A) για ένα ΝΠΑ A αν και μόνο αν η L είναι κανονική γλώσσα.

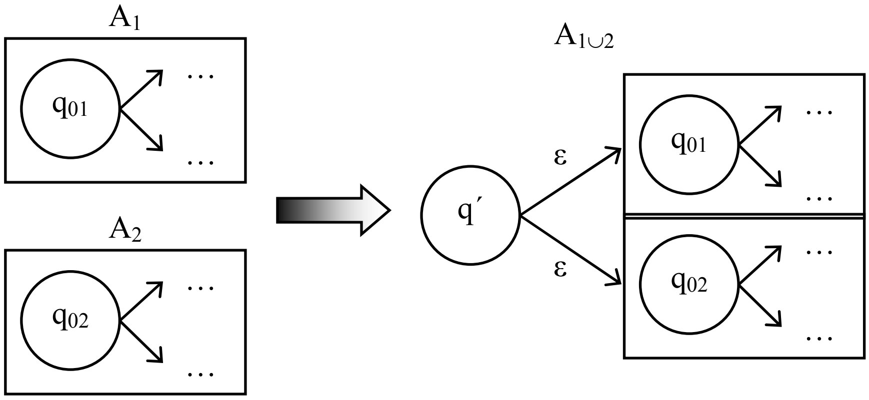
2.4 Ιδιότητες Κλειστότητας

Η κλάση των γλωσσών τύπου 3 είναι κλειστή ως προς:
- ✓ Ένωση
- ✓ Παράθεση (concatenation)
- ✓ Kleene Star
- ✓ Συμπλήρωση
- ✓ Τομή

Ένωση

Έστω $A_1=(Q_1, \Sigma, \delta_1, q_{01}, F_1)$, $A_2=(Q_2, \Sigma, \delta_2, q_{02}, F_2)$ δύο ΜΠΑ.
Υποθέτουμε ότι $Q_1\cap Q_2=\varnothing$ (διαφορετικά μετονομάζουμε τις καταστάσεις του A_2).
$A_{1\cup 2}=(Q_1\cup Q_2\cup\{q'\}, \Sigma, \delta, q', F_1\cup F_2)$ όπου:
- $q'\notin Q_1\cup Q_2$
- $\delta(q,\varepsilon)=\{q_{01}, q_{02}\}$ αν $q=q'$
- $\delta(q,\alpha)=\delta_1(q,\alpha)$ αν $q\in Q_1$
- $\delta(q,\alpha)=\delta_2(q,\alpha)$ αν $q\in Q_2$

Δηλαδή, ξεκινώντας από την αρχική κατάσταση q', το αυτόματο περνάει τυχαία στην αρχική κατάσταση είτε του A_1 είτε του A_2 και έπειτα μιμείται το A_1 ή το A_2.

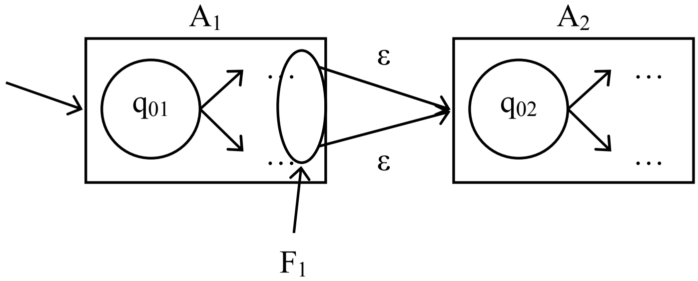


$$L(A_{1\cup 2})=L(A_1)\cup L(A_2)$$

➢ <u>Παρατήρηση</u>: χρησιμοποιήσαμε μεταβατικούς κανόνες της μορφής $\delta(q)=q'$ που κανονικά δεν επιτρέπονται σύμφωνα με τον ορισμό ΜΠΑ. Θα επανέλθουμε λίγο αργότερα σε αυτό το ζήτημα.

Παράθεση

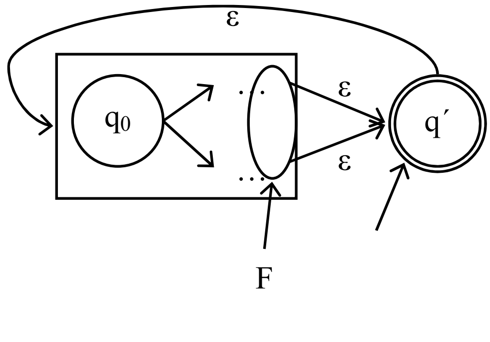
Έστω δύο ΜΠΑ $A_1=(Q_1, \Sigma, \delta_1, q_{01}, F_1)$, $A_2=(Q_2, \Sigma, \delta_2, q_{02}, F_2)$, $Q_1\cap Q_2=\varnothing$.
$A_{1\bullet 2}=(Q_1\cup Q_2, \Sigma, \delta, q_{01}, F_2)$ όπου $\delta=\delta_1\cup\delta_2\cup(F_1\times\{\varepsilon\}\times\{q_{02}\})$.
Δηλαδή ενώνουμε τις τελικές καταστάσεις του A_1 με την αρχική κατάσταση του A_2.



$$L(A_{1\bullet 2})=L(A_1)\bullet L(A_2)=\{w_1w_2 \mid w_1\in L(A_1), w_2\in L(A_2)\}$$

Kleene Star

Έστω $A=(Q, \Sigma, \delta, q_0, F)$ ένα ΜΠΑ.
Θέλουμε να κατασκευάσουμε ΜΠΑ A^* με $L(A^*)=(L(A))^*$.
$L(A^*)=(Q\cup\{q'\}, \Sigma, \delta', q', F\cup\{q'\})$, $q'\notin Q$.
$\delta'=\delta\cup(F\times\{\varepsilon\}\times\{q'\})\cup(\{q'\}\times\{\varepsilon\}\times\{q_0\})$.



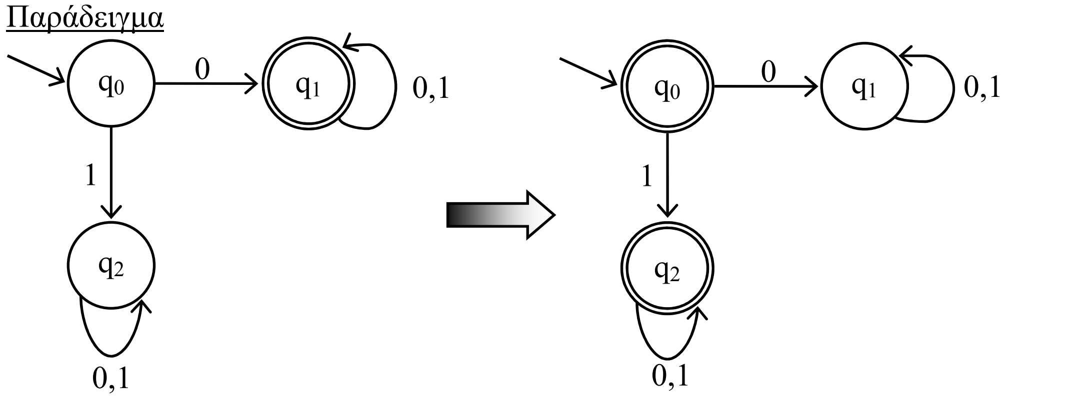
Συμπλήρωση

Έστω $A=(Q, \Sigma, \delta, q_0, F)$ ΝΠΑ.

Θέλουμε $L(A^{-})=(L(A))^{c}=\{w\in\Sigma^{*} \mid w\notin L(A)\}$.

$A^{-}=(Q, \Sigma, \delta, q_0, Q-F)$.

<u>Παράδειγμα</u>



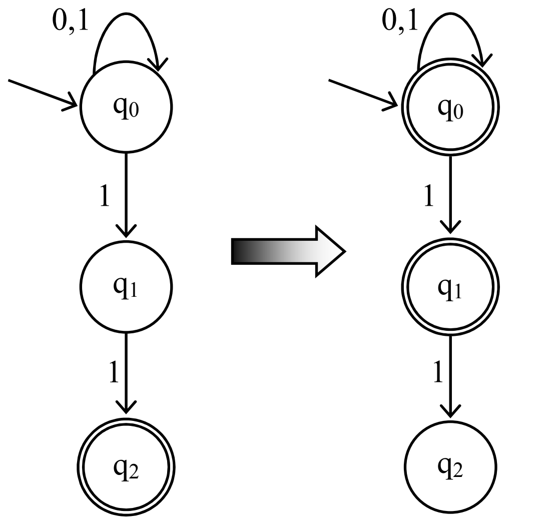
$L(A)=\{0w \mid w\in\{0,1\}^{*}\}$ $\qquad$ $L(A^{-})=\{1w \mid w\in\{0,1\}^{*}\}\cup\{\varepsilon\}=\{0,1\}^{*}-L(A)$

- Αλλά αυτή η κατασκευή *δεν* δουλεύει για ΜΠΑ!!!

<u>Παράδειγμα</u>

Στο διπλανό παράδειγμα ισχύει: $L(A')\neq L(A)^{c}$

(Το αυτόματο A είναι μη ντετερμινιστικό)



$L(A)=\{w11 \mid w\in\{0,1\}^{*}\}$ $\qquad$ $L(A')=\{0,1\}^{*}$!!

Τομή

Έστω L_1, L_2 κανονικές γλώσσες. Τότε ισχύει ότι:

$L_1\cap L_2=(L_1^{-}\cup L_2^{-})^{-}$.

Με βάση τα παραπάνω αποτελέσματα προκύπτει ότι η $L_1\cap L_2$ είναι κανονική γλώσσα.

- <u>Παρατήρηση</u>: έμμεση κατασκευή!

ΜΠΑ με ε-κανόνες

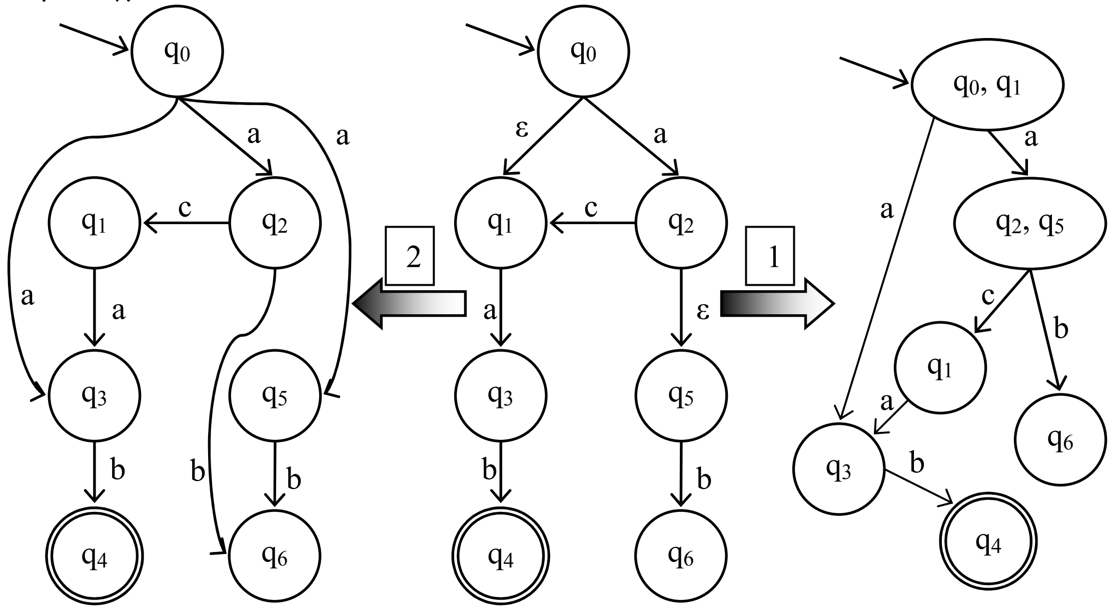
Σε μερικές περιπτώσεις χρησιμοποιήσαμε *ε-μεταβάσεις*, δηλαδή μεταβατικούς κανόνες της μορφής: $\delta(q,\varepsilon)=q'$.

Γενικά είναι μια ευκολία να έχουμε αυτή την δυνατότητα, αλλά δεν προσθέτει τίποτα καινούργιο. Δηλαδή για κάθε ΜΠΑ με ε-μεταβάσεις, υπάρχει ένα ισοδύναμο ΜΠΑ χωρίς ε-μεταβάσεις.

Υπάρχουν δύο τρόποι να το δει κανείς:

1. Χρησιμοποιώντας την ιδέα της κατασκευής ενός ΝΠΑ από ένα ΜΠΑ: με <u>σύνολα καταστάσεων</u>
2. Προσθέτοντας <u>νέες μεταβάσεις</u> που παρακάμπτουν τις ε-μεταβάσεις

<u>Παράδειγμα</u>



Συμπεραίνοντας Ιδιότητες Κανονικών Γλωσσών

<u>Θεώρημα 2.3</u> Δεδομένων αυτομάτων A, A_1, A_2, υπάρχει αλγοριθμικός τρόπος να εξακριβώσουμε τα εξής:

1. Αν $w \in L(A)$
2. Αν $L(A)=\varnothing$
3. Αν $L(A_1) \subseteq L(A_2)$
4. Αν $L(A_1)=L(A_2)$

<u>Σχέδιο Απόδειξης</u>

1. Εκτελούμε τις οδηγίες του ισοδύναμου ΝΠΑ.
2. Ελέγχουμε αν υπάρχει μία διαδρομή χωρίς κύκλους από το q_0 σε κάποιο $f \in F$.
3. $L(A_1) \subseteq L(A_2) \Leftrightarrow L(A_1) \cap (\Sigma^* - L(A_2)) = \varnothing$, οπότε με χρήση του 2 παραπάνω και των κατασκευών που έδειξαν ότι οι κανονικές γλώσσες είναι κλειστές ως προς τομή και συμπλήρωση προκύπτει το ζητούμενο.
4. $L(A_1)=L(A_2) \Leftrightarrow L(A_1) \subseteq L(A_2)$ και $L(A_2) \subseteq L(A_1)$.

2.5 Λήμμα Άντλησης για Κανονικές Γλώσσες

- Είναι ένας τρόπος να δείξουμε ότι μία γλώσσα *δεν* είναι κανονική.

Βασίζεται στην εξής απλή ιδέα:

- <u>Αρχή του Περιστερώνα</u>

n περιστερώνες m περιστέρια m>n	$\Rightarrow$	Τουλάχιστον δύο περιστέρια μοιράζονται τον ίδιο περιστερώνα.

<u>Θεώρημα 2.4</u> Για κάθε κανονική γλώσσα L υπάρχει ένας αριθμός $n \in \mathbb{N}$ έτσι ώστε κάθε λέξη $z \in L$ με $|z| \geq n$ μπορεί να τριχοτομηθεί σε $z=uvw$ με τις εξής ιδιότητες:

1. $|v| \geq 1$

2. $|uv| \leq n$
3. $uv^i w \in L, \forall i \geq 0$

<u>*Απόδειξη*</u>

Έστω L κανονική γλώσσα.
Έστω πεπερασμένο αυτόματο $A=(Q, \Sigma, \delta, q_0, F)$ με $L=L(A)$.
Ορίζουμε:
 $n:=\#Q$ (αριθμός καταστάσεων του A).
Έστω $w=x_1x_2\ldots x_m \in L$, $x_i \in \Sigma$, $m \geq n$.
Διαβάζοντας τα x_i ένα-ένα έχουμε μία σειρά $q_0,\ldots,q_m$ από καταστάσεις με:

- $q_i \in \delta(q_{i-1},x_i)$ για $i=1,2,\ldots,m$.
- $q_m \in F$, επειδή $w \in L$.

Αφού $m \geq n$ έπεται ότι υπάρχουν τουλάχιστον δύο ίδιες καταστάσεις μεταξύ των q_0, $q_1,\ldots, q_m$ (Αρχή του Περιστερώνα!).
Πιο συγκεκριμένα, μεταξύ των $q_0,q_1,\ldots,q_n$.
Έστω $q_j=q_k$, $0 \leq j<k \leq n$.
Ορίζουμε:

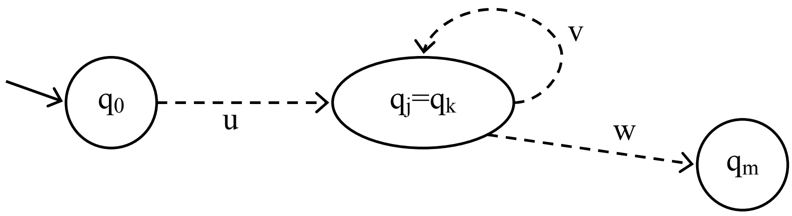
- $u=x_1\ldots x_j$
- $v=x_{j+1}\ldots x_k$
- $w=x_{k+1}\ldots x_m$

Προφανώς ισχύει:

- $\delta^*(q_0,u)=q_j$
- $\delta^*(q_j,v)=q_k$
- $\delta^*(q_k,w)=q_m$

Έχουμε:

- ✓ $j<k$, άρα: $|v| \geq 1$
- ✓ $j \leq n$, άρα: $|uv|=|x_1\ldots x_j|=j \leq n$
- ✓ $\delta^*(q_0,uw)=\delta^*(q_j,w)=\delta^*(q_k,w)=q_m \in F$
  $\delta^*(q_0,uv^kw)=\delta^*(q_j,v^kw)=\delta^*(q_k,v^{k-1}w)=$
  $=\delta^*(q_k,x^{k-2},w)=\ldots=\delta^*(q_k,w)=q_m \in F$



➢ Τι συμβαίνει διαισθητικά;

- Το A έχει έναν πεπερασμένο αριθμό n καταστάσεων. Αυτές είναι και η μόνη «μνήμη» του A.
- Αν διαβάσουμε μία μακρύτερη λέξη από n, τότε το A δεν μπορεί να διακρίνει εάν ένα κομμάτι της λέξης εμφανίζεται μία ή περισσότερες φορές!

<u>Παράδειγμα 1</u>

$L=\{a^kb^k \mid k \in \mathbb{N}\}$
<u>Διαισθητικά</u>: εάν $k>{}^n/_2$, το αυτόματο δεν μπορεί να θυμάται το ακριβές k!

<u>Επακριβής απόδειξη</u> ότι η L δεν είναι κανονική γλώσσα.

Έστω ότι είναι. Τότε το Λήμμα Άντλησης ισχύει. Έστω $n \in \mathbb{N}$ ο αριθμός που αναφέρεται στο λήμμα. Εξετάζουμε την λέξη $w=a^n b^n$. Προφανώς $|w| \geq n$, άρα υπάρχει $w=xyz$ με τις ιδιότητες που ορίζει το λήμμα.
Πώς μπορούν να είναι τα x, y, z;

a^n		b^n
$\Downarrow$		
u	v	w

$|uv| \leq n \Rightarrow v=a^k$
$|v|>1 \Rightarrow k>0$
Οπότε από Λήμμα παίρνουμε $uv^0w=uw \in L$. Αλλά $uv^0w=a^{n-k}b^k \notin L$, άτοπο.

<u>Παράδειγμα 2</u>
$L=\{w \in \{a,b\}^* \mid count_a(w)=count_b(w)\}$
<u>Διαισθητικά</u>: κανένα πεπερασμένο αυτόματο δεν μπορεί να αποθηκεύσει μία απεριόριστη διαφορά στον αριθμό των a, b!

<u>Επακριβής απόδειξη</u> ότι η L δεν είναι κανονική γλώσσα.
Έστω $w=a^n b^n \in L$. Όπως στο παράδειγμα 1 δείχνουμε ότι η L δεν είναι κανονική.

2.6 Κανονικές Εκφράσεις

- Οι κανονικές εκφράσεις είναι ένας τρίτος τρόπος προσδιορισμού των κανονικών γλωσσών.
- Περιγράφουν τον τρόπο δόμησης των κανονικών γλωσσών.
- Χρησιμοποιούνται στην πράξη (πχ στην αναζήτηση πληροφοριών σε λειτουργικά συστήματα – Unix).

Κανονικές Εκφράσεις
Έστω Σ ένα αλφάβητο με $\emptyset, \varepsilon, +, \bullet, {}^*, (,) \notin \Sigma$.
1. $\emptyset, \varepsilon \in REX(\Sigma)$
2. $\Sigma \subseteq REX(\Sigma)$
3. $a,b \in REX(\Sigma) \Rightarrow (a+b), (a \bullet b), (a^*) \in REX(\Sigma)$
4. Τίποτα δεν είναι κανονική έκφραση, εκτός αν προκύπτει από τα 1-3

Η Γλώσσα μίας Κανονικής Έκφρασης
1. $L(\emptyset)=\emptyset$, $L(\varepsilon)=\{\varepsilon\}$
2. $L(a)=\{a\}\ \forall a \in \Sigma$
3. $L(a+b)=L(a) \cup L(b)$
   $L(a \bullet b)=L(a)L(b)=\{w_1 w_2 \mid w_1 \in L(a), w_2 \in L(b)\}$
   $L(a^*)=L(a)^*=\{w_1 w_2 \ldots w_n \mid w_1, w_2, \ldots, w_n \in L(a), n \in \mathbb{N}\}$

Πολλές παρενθέσεις μπορούν να παραληφθούν, θεωρώντας το * ισχυρότερο, μετά το • και πιο αδύνατο το +. Επίσης το • παραλείπεται (όπως ο πολλαπλασιασμός στην αριθμητική).
Πχ: $(((a \bullet b)^*) \bullet (c \bullet (c^*)))=(ab)^* cc^*$

<u>Θεώρημα 2.5</u> Η L είναι κανονική γλώσσα αν και μόνο αν $L=L(a)$ για μία κανονική έκφραση a.