ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΜΟΡΦΟΛΟΓΙΑ

- ΚΑΝΟΝΙΚΕΣ ΕΚΦΡΑΣΕΙΣ
- ΑΥΤΟΜΑΤΑ ΠΕΠΕΡΑΣΜΕΝΩΝ ΚΑΤ/ΣΕΩΝ
- ΜΕΤΑΤΡΟΠΕΙΣ ΠΕΠΕΡΑΣΜΕΝΩΝ ΚΑΤ/ΣΕΩΝ

Τι είναι η Υπολογιστική Μορφολογία;

Η μελέτη της αυτόματης αντιστοίχισης μιας λέξης στις μορφολογικές πληροφορίες που τη χαρακτηρίζουν και το αντίστροφο

- Μορφολογική ανάλυση (parsing)

  cats: cat + N + PL
- Μορφολογική παραγωγή (generation)

  cat + N + PL: *cats*
- Χρήσιμα εργαλεία για την επίλυση του προβλήματος είναι
  - τα αυτόματα και
  - οι μετατροπείς πεπερασμένων καταστάσεων

Γιατί είναι σημαντική;

- Ως προεπεξεργασία σε συστήματα επεξεργασίας φυσικής γλώσσας
  - π.χ. εξαγωγή πληροφορίας, μηχανική μετάφραση κτλ.
- Εφαρμογές επεξεργασίας κειμένου
  - π.χ. έλεγχος ορθογραφίας
- Ανάκτηση πληροφορίας
  - π.χ. stemming
- Συστήματα επεξεργασίας ομιλίας
- Χρήση ηλεκτρονικών λεξικών
  - π.χ. λημματοποίηση

Μορφολογική ταξινόμηση γλωσσών

- Απομονωτικές γλώσσες
- Συνθετικές Γλώσσες
  - Συγκολλητικές γλώσσες
  - Πολυσυνθετικές γλώσσες
  - Κλιτές γλώσσες

Απομονωτικές γλώσσες (isolating)

- Είναι οι γλώσσες που κάθε λέξη αποτελείται από ένα μόρφημα (π.χ. Βιετναμέζικα, Κινέζικα).
- Σε αντίθεση με τις συνθετικές γλώσσες (synthetic) όπου μια λέξη μπορεί να σχηματίζεται από περισσότερα μορφήματα.
- **Μόρφημα** είναι η μικρότερη γλωσσολογική μονάδα που φέρει μορφολογική πληροφορία.
- Σε αυτές τις γλώσσες κάποιες μορφολογικές πληροφορίες μιας λέξης δεν δηλώνονται εμφανώς (π.χ. στα κινέζικα ο χρόνος και ο αριθμός)

	gou	*bu*	*ai*	*chi qingcai*
(the)	*dog/s*	*do/does/did*	*not*	*like eat vegetables*

Πολυσυνθετικές γλώσσες (polysynthetic)

- Είναι οι γλώσσες που κάθε λέξη αποτελείται από πολλά μορφήματα και αντιστοιχούν σε ολόκληρη πρόταση (γλώσσες Εσκιμώων, όπως η γλώσσα Yupik κεντρικής Αλάσκας)

  qaya:liyu:lu:ni

  'ήταν τέλειος (-yu-) στο να φτιάχνει (-li-) καγιάκ (qaya:-)'

Συγκολλητικές γλώσσες (agglutinative)

- Είναι οι γλώσσες που οι λέξεις τους σχηματίζονται από πολλά μορφήματα με ευδιάκριτα όρια (Τούρκικα, Ουγγρικά, Σουαχίλι)
- Κάθε μόρφημα εμπεριέχει μια μόνο μορφολογική πληροφορία

*Cop+luk+**ler**+imiz+de+ki+**ler**+den+mi+y+di*

σκουπίδια+AFF+**PL**+1P/PL+LOC+REL+**PL**+ABL+INT+AUX+PAST

'ήταν από εκείνα που ήταν στους σκουπιδοντενεκέδες μας;'

Κλιτές γλώσσες (inflectional)

- Είναι οι γλώσσες που χρησιμοποιούν μεγάλο αριθμό μορφημάτων.
- Ένα μόρφημα εκφράζει ταυτόχρονα πολλές μορφολογικές κατηγορίες. (Ελληνικά, Λατινογενείς γλώσσες)

  παίζ-ουν: 3ο+PL+ACT+PRES/PAST

  vogli-o: 1ο+SNG+ACT+PRES/PAST

Λήμμα (Lemma)

- **Λήμμα** είναι ένα σύνολο μορφών της ίδιας λέξης. Οι μορφές έχουν
  - την ίδια ρίζα,
  - το ίδιο μέρος του λόγου,
  - την ίδια έννοια.
- Παράδειγμα: Οι παρακάτω λέξεις έχουν το ίδιο λήμμα (*παίζω*):

  παίζω, παίζεις, παίζει, παίζουμε, έπαιξε…

Παραγωγή (derivation)

- Ο σχηματισμός λέξεων διαφορετικών κατηγοριών από το ίδιο λήμμα

establish (V)

establish+*ment* (N)

establish+*ment*+*ary* (Adj)

establish+*ment*+*ari*+*an* (N)

establish+*ment*+*ari*+*an*+*ism* (N)

dis+*establish*+*ment*+*ari*+*an*+*ism* (N)

anti+*dis*+*establish*+*ment*+*ari*+*an*+*ism* (N)

Προσφυματοποίηση (affixation)

- Και η κλίση και η παραγωγή στηρίζονται στην προσφυματοποίηση (εισαγωγή προθέματος, κατάληξης, ενθέματος).
- Πρόθεμα (prefix): ***ξε-****πλένω*
- Ένθεμα (infix): *δι-**έ-**σχισα*
- Κατάληξη (suffix): *κατεργαρ-**άκος***

Μόρφημα

- Μόρφημα: ΓΕΝΙΚΗ
- Μορφές:
  - Διαφορετικές μορφές ανάλογα με το λήμμα και τα χαρακτηριστικά του

    {-ας} / {-ου} / {-ους} / {-α} ...

    χώρ-ας *λόγ-ου* *λάθ-ους* *πατέρ-α*
  - Διαφορετικές μορφές για το ίδιο λήμμα

    γράφ {-ονταν} / {-όντανε} / {-όντουσαν}

Κανονικές εκφράσεις (Regular Expressions)

www.regexpal.com

- Οι **κανονικές εκφράσεις** είναι αλγεβρικές εκφράσεις που αντιπροσωπεύουν ένα πρότυπο (pattern) ακολουθίας συμβολοσειρών (strings).
- Χρησιμοποιούνται
  - για την αναζήτηση συμβολοσειρών σε κείμενο
  - για τον τυπικό ορισμό μιας γλώσσας
- Παραδείγματα
  - Απλή ακολουθία χαρακτήρων

    /δώρο/ *Δώσε μου το δώρο μου.*

    /!/ *Δεν είσαι καλά!*

    /αδερφό μου/ *Μίλησα στον αδερφό μου χτές.*
  - Πολλαπλή ταυτοποίηση

    /[Δδ]ώρο/ *δώρο, Δώρο*

    /[0123456789]/ *όλα τα ψηφία*

    /[3-5]/ *3,4,5* *(τα ψηφία από το 3 ως το 5)*

ΚΕ: Παραδείγματα (συν)

- Άρνηση

  /[^α]/ *Λες σαχλαμάρες*

  /[^0-9]/ *ταυτοποιείται με όλα εκτός από ψηφία*

- Προαιρετικότητα

  /πατέρας?/ *πατέρα, πατέρας* *(ο χαρ/ρας ακριβώς πριν από το ? μπορεί να υπάρχει ή όχι)*

- Επανάληψη
  - ο *χαρ/ρας ακριβώς πριν από το * μπορεί να υπάρχει καμία ή περισσότερες φορές*

    /μπα*/ *μπ, μπα, μπαα, μπααα, μπαααα, ...*

    /[αβ]*/ *αααα, βββββββββ, αβαβαβαβ, κενό, ...*
  - (ο *χαρ/ρας ακριβώς πριν από το ! μπορεί να υπάρχει μία ή περισσότερες φορές*)

    /μπα!/ *μπα, μπαα, μπααα, μπαααα, ...*
  - Το ! πολλές φορές συναντάται ως +

ΚΕ: Παραδείγματα (συν)

- Χαρακτήρας μπαλαντέρ (.)

  /φ.λο/ *φίλο*, *φύλο*

- Αρχή γραμμής κειμένου

  /^Το/ ... . *Το σπίτι μου...*

- Τέλος γραμμής κειμένου

  /ωραίο\.$/ *....είναι ωραίο. Επίσης...*

- Διάζευξη

  /σκύλος|γάτα/ *Αυτός είναι ο σκύλος και αυτή είναι η γάτα μου.*

  /σκύλο(ς|ι)/ *σκύλος*, *σκύλοι*

- Αρχή-τέλος λέξης

  /\bστο\b/ *Με αποστομώνει στο τέλος.*

ΚΕ: Προτεραιότητα τελεστών

παρενθέσεις	()
μετρητές	? + *
αρχή/ τέλος γραμμής/λέξης	^ $ \b
διάζευξη	|

Παραδείγματα

- Έστω ότι θέλω να φτιάξω μια ΚΕ που να βρίσκει το άρθρο *το* μέσα σε ένα κείμενο.

  /το/ Δεν βρίσκει το *Το*.

  /[Ττ]ο/ Βρίσκει και τα: *στο, απότομο κλπ.*

  /\b[Ττ]ο\b/

- [^A-Za-z][A-Za-z]+ing[^A-Za-z]
- [^A-Za-z][Tt]he[^A-Za-z]

- Για κάθε μια από τις ακόλουθες ΚΕ δώστε δυο συμβολοσειρές που ανήκουν στην γλώσσα που αναπαριστά η ΚΕ και δυο που δεν ανήκουν.

  1) *a*b**

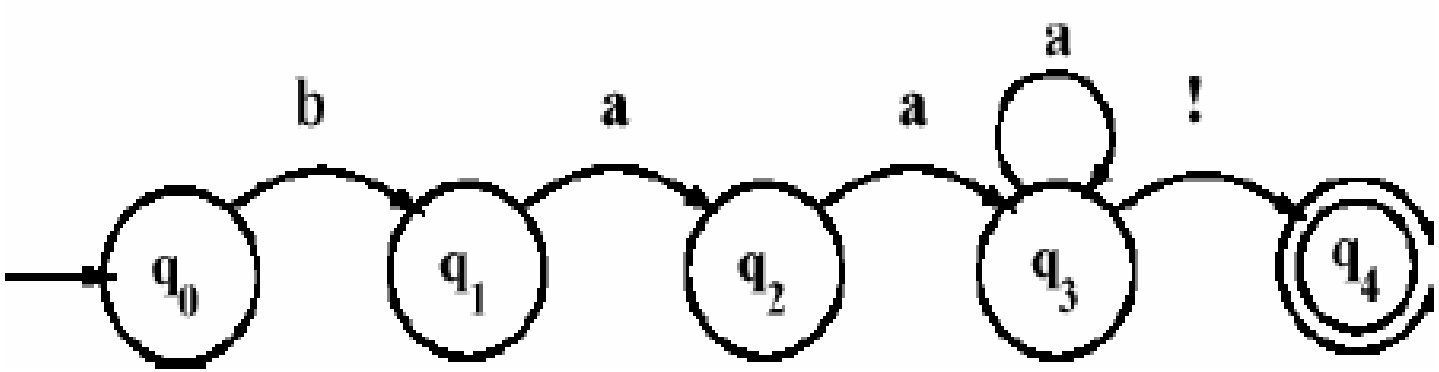
  2) *a(ba)*a*

Αυτόματα πεπερασμένων καταστάσεων (Finite State Automata - FSAs)

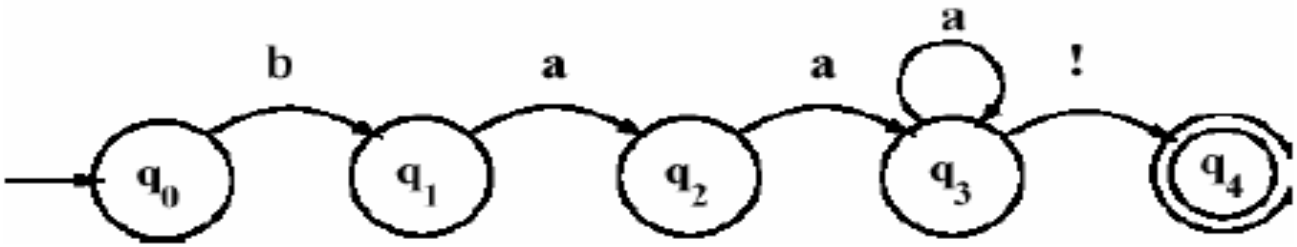
- Ένα ΑΠΚ είναι ένας γράφος που αποτελείται από
  - κόμβους (καταστάσεις) και
  - κατευθυνόμενα βέλη (μεταβάσεις), που αντιστοιχούν στα σύμβολα ενός αλφαβήτου
- Μια κατάσταση ορίζεται ως αρχική
  - Συμβολίζεται με ένα εισερχόμενο βέλος
- Μια ή περισσότερες καταστάσεις ορίζονται ως τελικές ή αποδεκτές
  - Συμβολίζονται με διπλό κύκλο

ΑΠΚ: Παράδειγμα

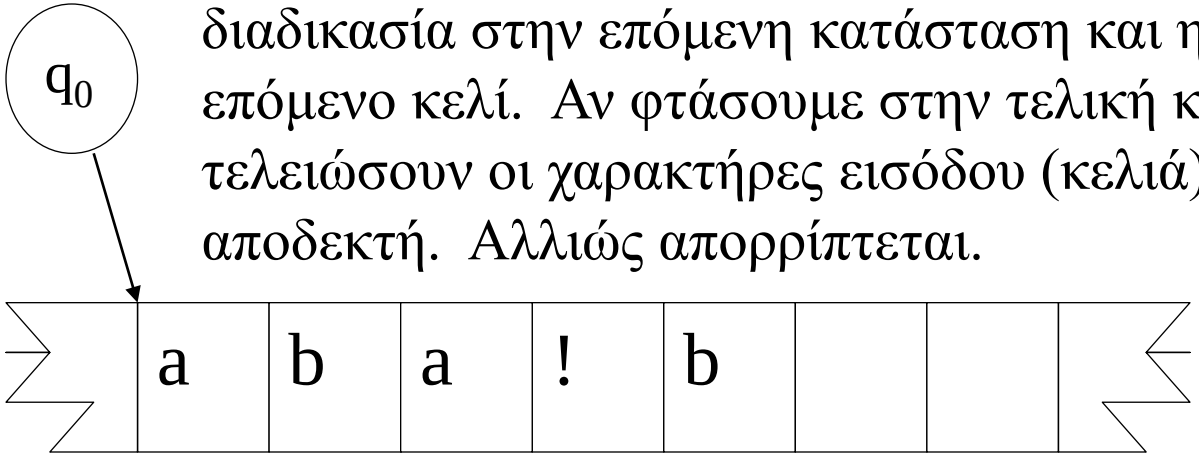
- baa! baaa! baaaa!
- Αυτό το αυτόματο είναι ισοδύναμο με την κανονική έκφραση /baaa*\!/



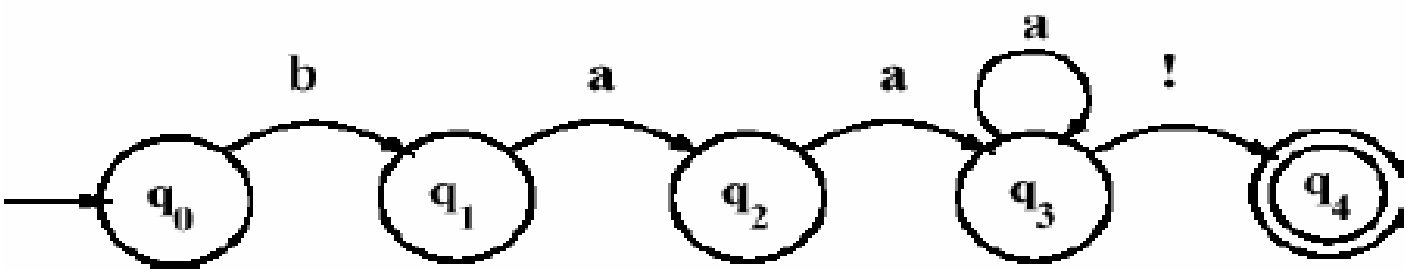
Αναγνώριση αλφαριθμητικών με ΑΠΚ: Μη αποδεκτή λέξη

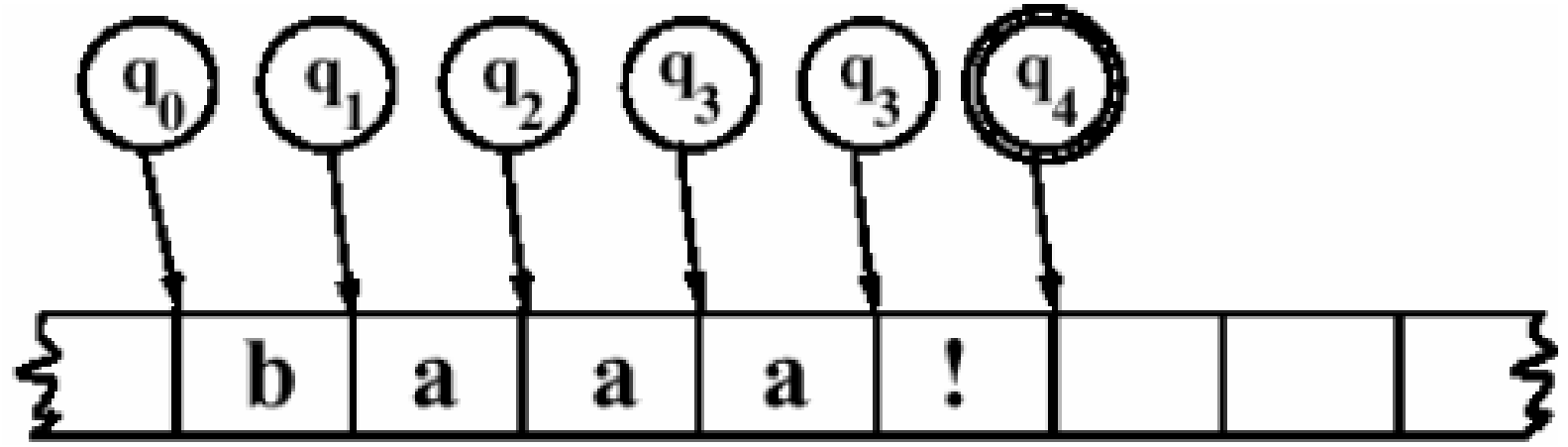


Σε κάθε κελί της ταινίας υπάρχει και ένας χαρακτήρας.
Η αρχική κατάσταση δείχνει στην αρχή του πρώτου κελιού.
Αν ο χαρακτήρας εισόδου ταιριάζει με τον χαρακτήρα στο βέλος της αντίστοιχης κατάστασης του αυτομάτου, περνάει η διαδικασία στην επόμενη κατάσταση και η ταινία προχωράει στο επόμενο κελί. Αν φτάσουμε στην τελική κατάσταση όταν μας τελειώσουν οι χαρακτήρες εισόδου (κελιά), τότε η λέξη είναι αποδεκτή. Αλλιώς απορρίπτεται.

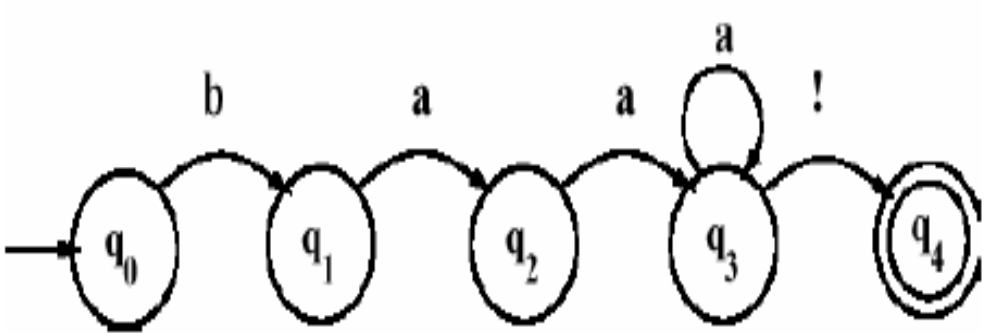


Αναγνώριση αλφαριθμητικών με ΑΠΚ: Αποδεκτή λέξη





Πίνακας Μετάβασης Καταστάσεων



	Input		
State	b	a	!
0	1	0	0
1	0	2	0
2	0	3	0
3	0	3	4
4:	0	0	0

Τυπικός Ορισμός ενός ΑΠΚ

- ***Q***: ένα πεπερασμένο σύνολο καταστάσεων (q_0, q_1, …)
- ***Σ***: ένα πεπερασμένο αλφάβητο συμβόλων εισόδου
- q_0: η αρχική κατάσταση
- ***F***: οι τελικές καταστάσεις (υποσύνολο του Q)
- ***δ(q,i)***: η συνάρτηση μετάβασης καταστάσεων. Δεδομένης μιας κατάστασης q και ενός συμβόλου εισόδου i, επιστρέφει μια νέα κατάσταση q'

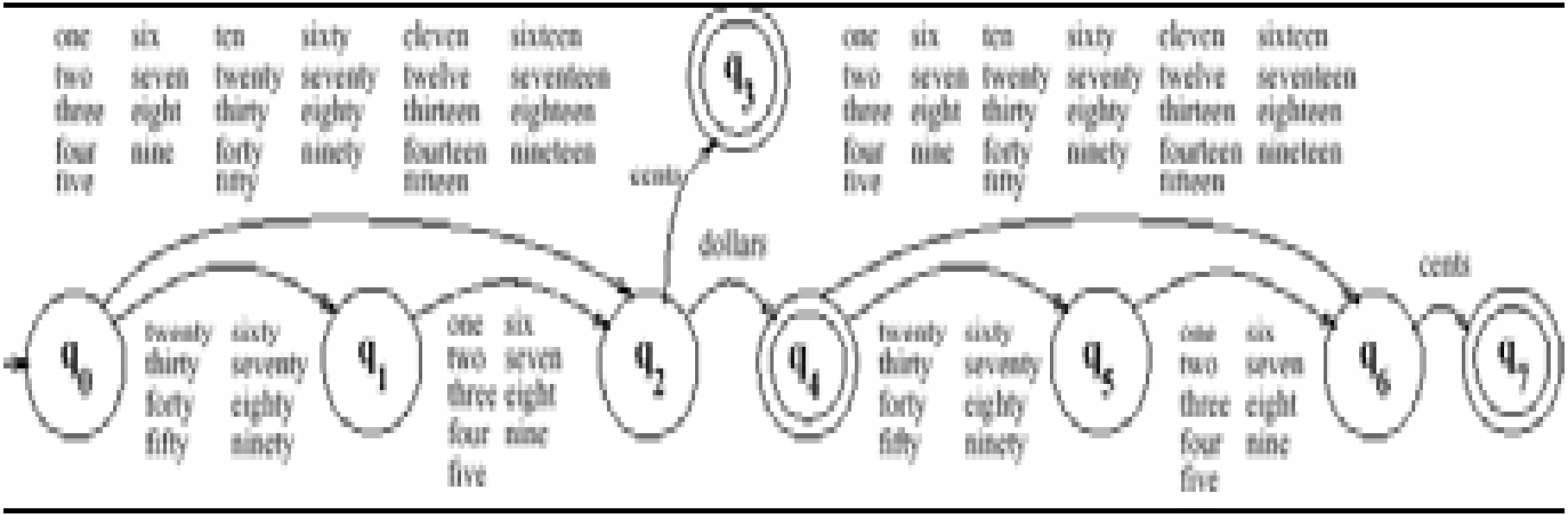
Τυπικές Γλώσσες (Formal Languages)

- Η **τυπική γλώσσα** είναι ένα σύνολο από αλφαριθμητικά (λέξεις).
- Κάθε **λέξη** απαρτίζεται από σύμβολα.
- Το σύνολο των συμβόλων είναι το **αλφάβητο** της γλώσσας.
- Το αλφάβητο του προηγούμενου παραδείγματος είναι το σύνολο Σ={a, b, !}
- Η τυπική γλώσα του προηγούμενου παραδείγματος είναι άπειρο σύνολο

  L={baa!, baaa!, baaaa!, baaaaa!, …}

Παράδειγμα ΑΠΚ

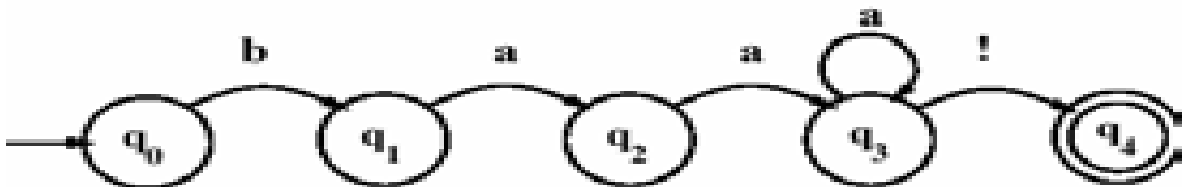
Το αλφάβητο μπορεί να είναι και υψηλότερου επιπέδου, να αποτελείται από λέξεις αντί για σύμβολα.



Τα αγγλικά χρηματικά ποσά (μέχρι το ποσό $99.99)

Ντετερμινιστικά-Μη ντετερμινιστικά ΑΠΚ

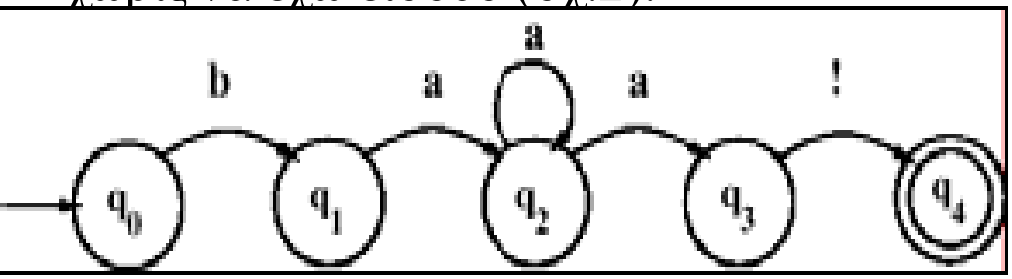
Deterministic FSA (DFSA): Η αναγνώριση μιας συμβολοσειράς δεν υπόκειται σε επιλογές.



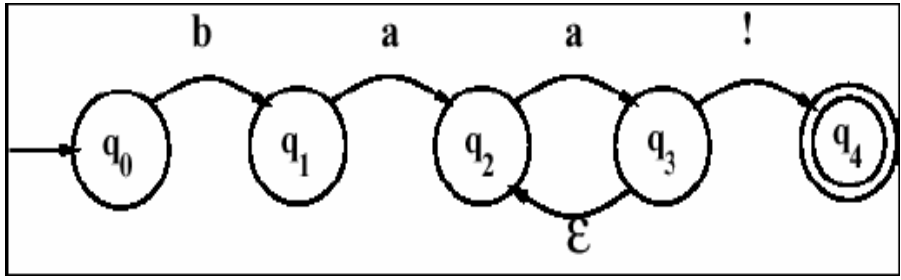
Non-Deterministic FSA (NFSA):

• Όταν από μία κατάσταση και με την ίδια είσοδο, το FSA έχει περισσότερες από μια επιλογές για την κατάσταση που θα μετακινηθεί (σχ.1).

• Όταν έχω μεταβάσεις ε (ε-transitions). Οι μεταβάσεις αυτές (κενά τόξα) με μεταφέρουν από μια κατάσταση σε άλλη χωρίς να λαμβάνεται υπόψη η είσοδος ή χωρίς να έχω είσοδο (σχ.2).



Σχήμα 1



Σχήμα 2

Πίνακας Μετάβασης NFSA

- Προστίθεται μια επιπλέον στήλη για το κενό τόξο
- Οι μεταβάσεις είναι σύνολα καταστάσεων

	Input			
State	**b**	**a**	**!**	**ε**
0	1	Ø	Ø	Ø
1	Ø	2	Ø	Ø
2	Ø	2,3	Ø	Ø
3	Ø	Ø	4	2
4:	Ø	Ø	Ø	Ø

Αναγνώριση με NFSA

- Μια λέξη μπορεί να απορρίπτεται λανθασμένα από ένα NFSA επειδή επιλέχτηκε το λάθος τόξο. Αυτό αντιμετωπίζεται
  - Με σήμανση του κελιού της ταινίας στο οποίο έχουμε φτάσει μέχρι την στιγμή που εμφανίζεται η μη ντετερμινιστική επιλογή. Έτσι, αν αποδειχθεί ότι κάναμε λάθος επιλογή μπορούμε να γυρίσουμε πίσω και να επιλέξουμε διαφορετικό μονοπάτι.
  - Με έλεγχο της συνέχειας της εισόδου. Κοιτάω από πριν πώς εξελίσσεται η είσοδος και επιλέγω ανάλογα.
  - Με παράλληλη εξέταση εναλλακτικών μονοπατιών.

Αναγνώριση με NFSA

- Η αναγνώριση συμβολοσειρών με NFSA μπορεί να θεωρηθεί ως αναζήτηση σε ένα χώρο καταστάσεων
- Η σειρά με την οποία εξετάζονται οι καταστάσεις (η απόφαση για το ποιο μονοπάτι θα ακολουθηθεί) επηρεάζει την επίδοση
- Αναζήτηση σε βάθος ή αναζήτηση σε πλάτος
- Για μεγάλους χώρους καταστάσεων είναι απαραίτητο να χρησιμοποιηθούν πιο προηγμένες τεχνικές αναζήτησης (Δυναμικός προγραμματισμός ή A*)

Μορφολογική Ανάλυση-Μορφολογική Αναγνώριση

- **Μορφολογική Αναγνώριση:** Αποδέχεται ή απορρίπτει λέξεις
  - goose
    - gooses x
    - geese √
- **Μορφολογική Ανάλυση (Parsing):** Εξάγει μορφολογική πληροφορία σχετικά με μία λέξη (λήμμα, μορφολογικά χαρακτηριστικά)
  - geese: goose + N + PL
  - cats: cat + N + PL
  - ground: ground +N +SG
- **Μορφολογική Σύνθεση (Generation/Synthesis):** Σχηματίζει μια λέξη από το λήμμα και τα μορφολογικά της χαρακτηριστικά (η αντίθετη διαδικασία από την ανάλυση)

Μορφολογικός Αναλυτής

- Ένας μορφολογικός αναλυτής αποτελείται από
  - Λεξικό: μία λίστα από λήμματα ή καταλήξεις μιας γλώσσας, μαζί με τη βασική πληροφορία που τα χαρακτηρίζουν

    Π.χ. dog: reg-noun

    goose: irreg-sg-noun

    geese: irreg-pl-noun

    -s: plural-suffix
  - Μορφοτακτικοί κανόνες: ένα μοντέλο διάταξης των μορφημάτων που καθορίζει ποιες κατηγορίες μορφημάτων ακολουθούνται από άλλες κατηγορίες (π.χ. ότι το μόρφημα του πληθυντικού στα αγγλικά ακολουθεί το ουσιαστικό)
  - Ορθογραφικοί κανόνες: μοντελοποιούν ορθογραφικές αλλαγές που συμβαίνουν στις λέξεις μίας γλώσσας

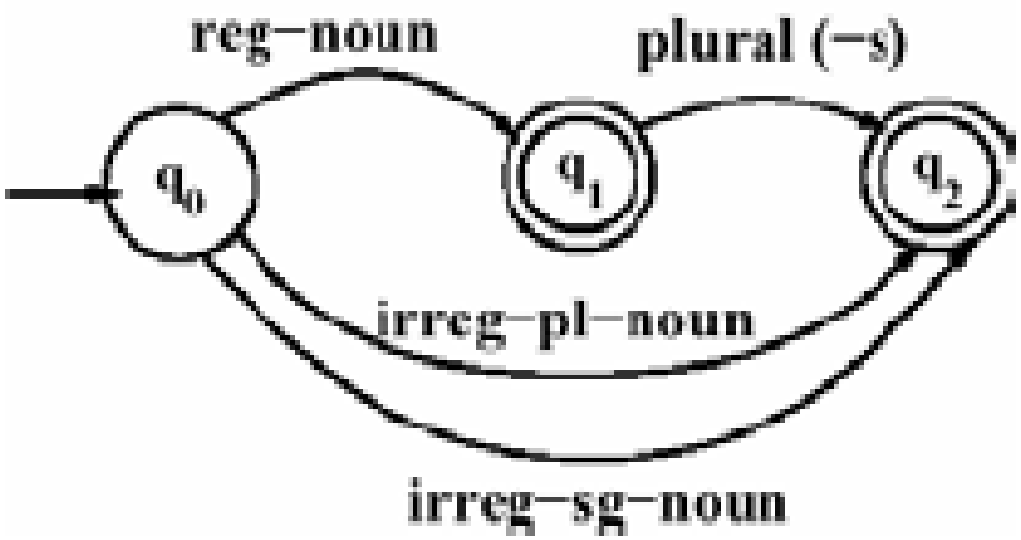
    (π.χ. city+s -> cities)

Λεξικό

- Στην πιο απλή περίπτωση είναι ένα σύνολο από όλες τις λέξεις της γλώσσας:
  - a, AAA, AA, Aachen, aardvark, aardwolf...
- Δεν είναι πρακτικό να αποθηκεύσουμε όλες τις δυνατές λέξεις μιας γλώσσας
  - Για μερικές γλώσσες (Φιλανδικά, Τουρκικά) αυτό είναι αδύνατο
  - Συνήθως αποθηκεύονται μόνο οι ρίζες και οι καταλήξεις
- Στην ιδεατή περίπτωση όλες οι πιθανές λέξεις (π.χ. ακρωνύμια, κύρια ονόματα κτλ.) πρέπει να βρίσκονται στο λεξικό

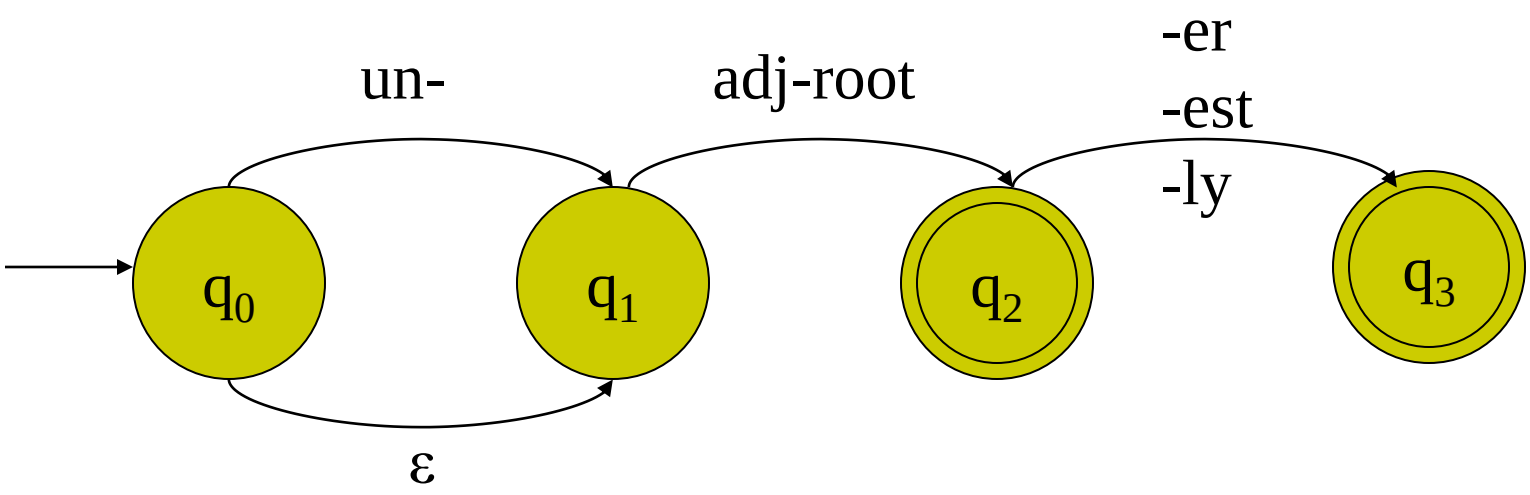
Μορφοτακτικοί κανόνες

- Συνήθως αναπαριστάνονται με ΑΠΚ
- Π.χ. Ένα απλό FSA για το σχηματισμό πληθυντικού αριθμού στα Αγγλικά:



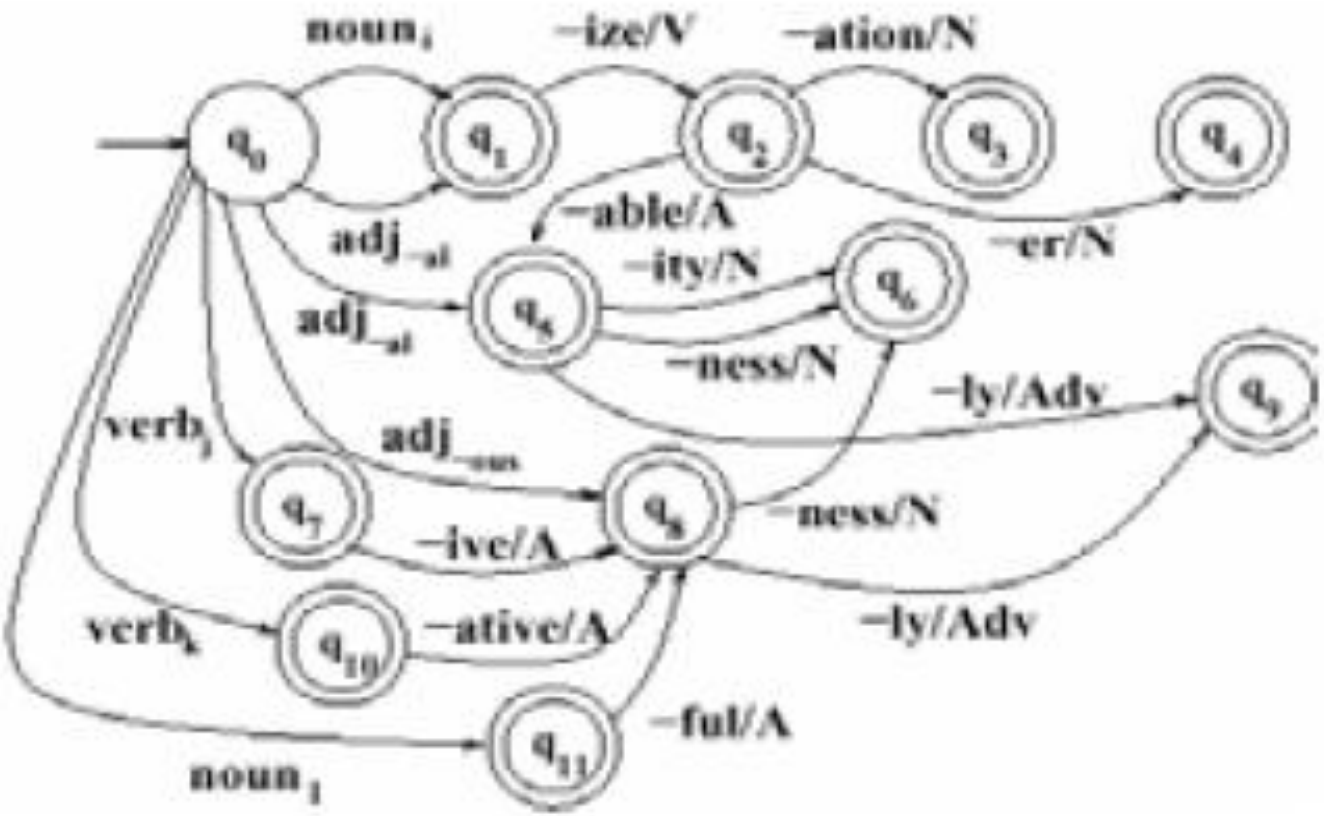
Μορφολογία αγγλικών επιθέτων

- clear, clearer, clearest, clearly, unclear, unclearly
- cool, cooler, coolest, coolly
- big, bigger, biggest, **unbig, bigly**
- red, redder, reddest, **redly**
- real, really, **realer**, **realest**

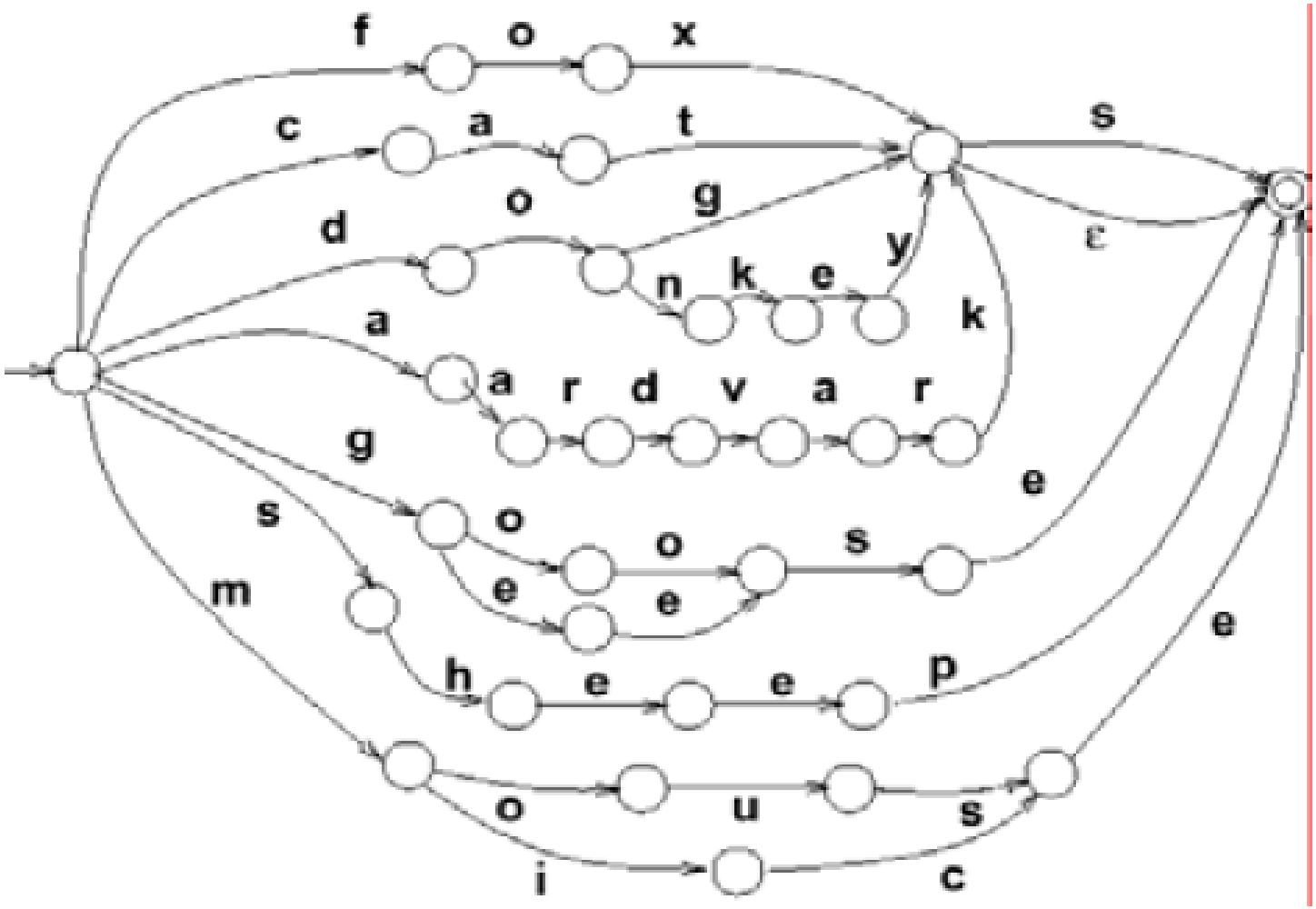


Πολυπλοκότητα

- Ένα ΑΠΚ για περιγραφή τμήματος της αγγλικής παραγωγικής μορφολογίας

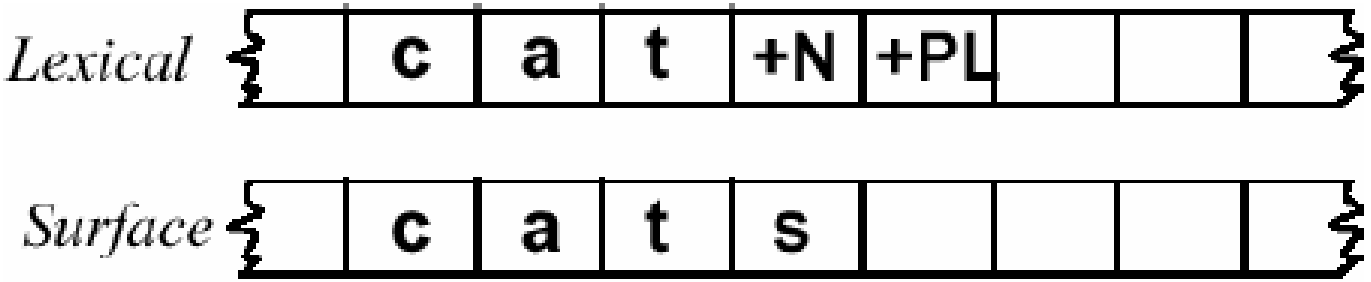


Μορφολογική Αναγνώριση με ΑΠΚ



Μορφολογική Ανάλυση

- Μορφολογία 2 επιπέδων (Two-level Morphology)
- **Επιφανειακό επίπεδο λέξης (surface level):** η λέξη όπως αυτή εμφανίζεται σε ένα κείμενο
- **Λεξικολογικό επίπεδο λέξης (lexical level)**: μια ακολουθία (συνένωση) μορφημάτων που σχηματίζουν την λέξη
- Η μορφολογία 2 επιπέδων αναπαριστά μια λέξη σαν μια αντιστοιχία ανάμεσα στο λεξικολογικό και το επιφανειακό επίπεδο



Μετατροπέας Πεπερασμένων Καταστάσεων (Finite State Transducer - FST)

- Ο ΜΠΚ πραγματοποιεί την αντιστοιχία μεταξύ των 2 επιπέδων
- Ο ΜΠΚ είναι ένα αυτόματο με 2 ταινίες που αναγνωρίζει ή παράγει **ζευγάρια** συμβολοσειρών.
- Ο ΜΠΚ είναι μια μηχανή που διαβάζει μια συμβολοσειρά και παράγει μια άλλη.

Μετατροπέας Πεπερασμένων Καταστάσεων

- Σαν μηχανή αναγνώρισης, ο ΜΠΚ παίρνει σαν είσοδο δύο συμβολοσειρές και βγάζει ως έξοδο
  - Αποδοχή, εάν το ζεύγος των συμβολοσειρών εμπεριέχεται στην γλώσσα, και
  - Απόρριψη, εάν δεν εμπεριέχεται

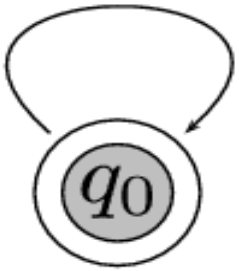
- Σαν μηχανή μετάφρασης, ο ΜΠΚ παίρνει μια συμβολοσειρά και παράγει μια άλλη συμβολοσειρά

Τυπικός ορισμός ενός ΜΠΚ

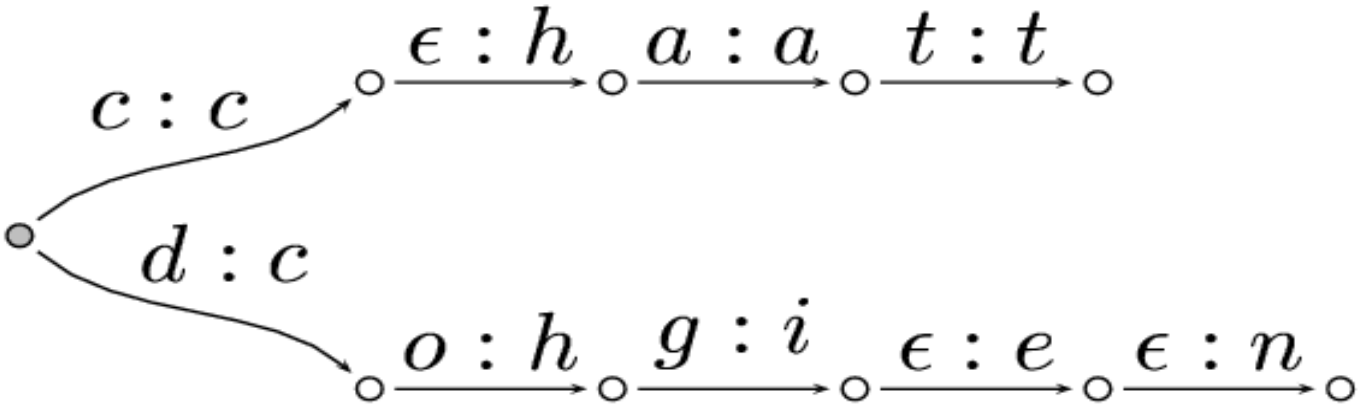
- **Q**: ένα πεπερασμένο σύνολο καταστάσεων (q_0, q_1, …)
- **Σ**: ένα πεπερασμένο αλφάβητο ζευγών συμβόλων *i:o*, όπου
  - *i* είναι ένα σύμβολο από το αλφάβητο εισόδου και
  - *o* ένα σύμβολο από το αλφάβητο εξόδου
    (το ε μπορεί να ανήκει και στα δύο αλφάβητα)
  - Π.χ. Σ= {a:a, b:b, !:!, a:!, a:b, b:a, a:ε, ε:!}
- **q_0**: η κατάσταση εκκίνησης
- **F**: οι τελικές καταστάσεις (υποσύνολο του Q)
- **δ(q,*i:o*)**: η συνάρτηση μετάβασης. Δεδομένης μίας κατάστασης q και μίας εισόδου *i*, επιστρέφει μία νέα κατάσταση q’.

Example: The uppercase transducer

$$a : A, b : B, c : C, \ldots$$



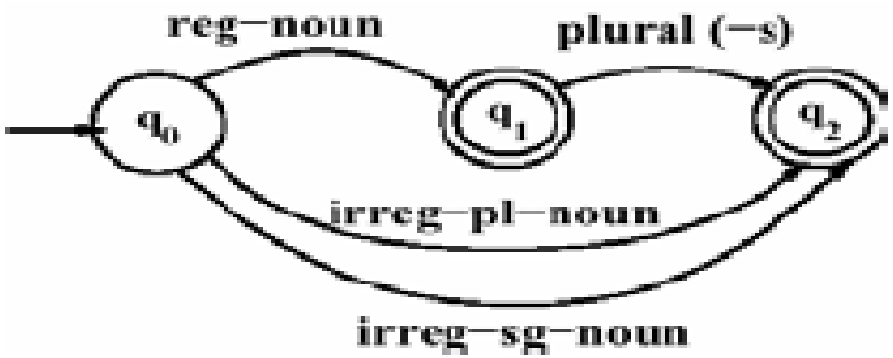
Example: English-to-French



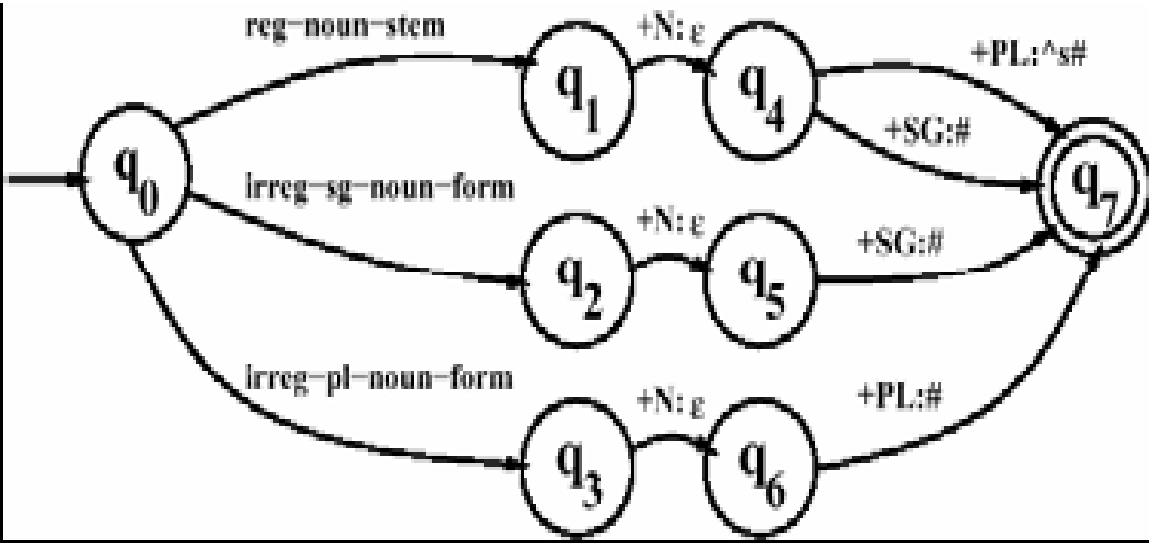
ΜΠΚ και Μορφολογία 2 επιπέδων

- Είναι βολικό να θεωρήσουμε ότι ένας ΜΠΚ έχει δύο ταινίες
  - **Λεξικολογική ταινία (lexical tape)**
  - **Επιφανειακή ταινία (surface tape)**
- Κάθε ζεύγος συμβόλων a:b εκφράζει πώς αντιστοιχίζεται ένα σύμβολο της μίας ταινίας σε κάποιο σύμβολο της άλλης
- Στην λεξικολογική ταινία υπάρχουν τα σύμβολα του αριστερού μέρους των ζευγών a:b
- Στην επιφανειακή ταινία υπάρχουν τα σύμβολα του δεξιού μέρους των ζευγών a:b
- Σύμβολα όπως το *a:a* καλούνται **default pairs** και συνήθως αναπαρίστανται απλά ως *a*.

Από ΑΠΚ σε ΜΠΚ



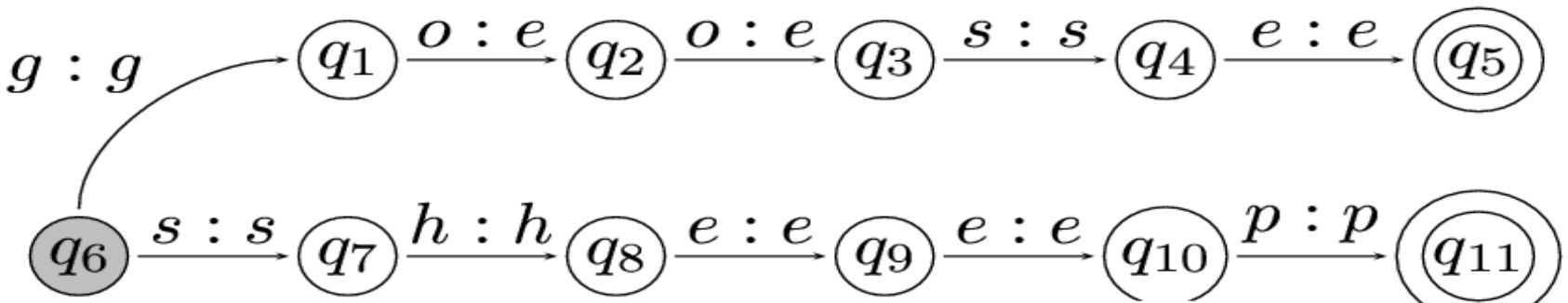
Το ΑΠΚ για τον σχηματισμό του πληθυντικού στα Αγγλικά



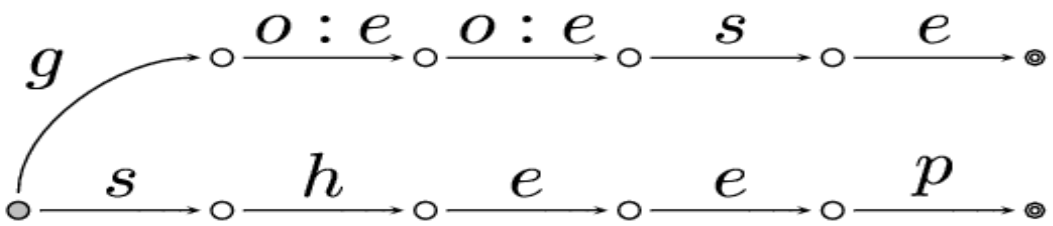
Η εξέλιξη του ΑΠΚ σε ΜΠΚ με την προσθήκη των μορφ/κών χαρακτηριστικών +SG, +PL, +N.
Τα χαρ/κά αυτά σχηματίζουν ζευγάρια με τα σύμβολα
ε (κενή συμβολοσειρά)
^ (όριο μορφήματος)
# (όριο λέξης)
γιατί δεν εμφανίζονται στην ταινία εξόδου.

Το λεξικό του ΜΠΚ

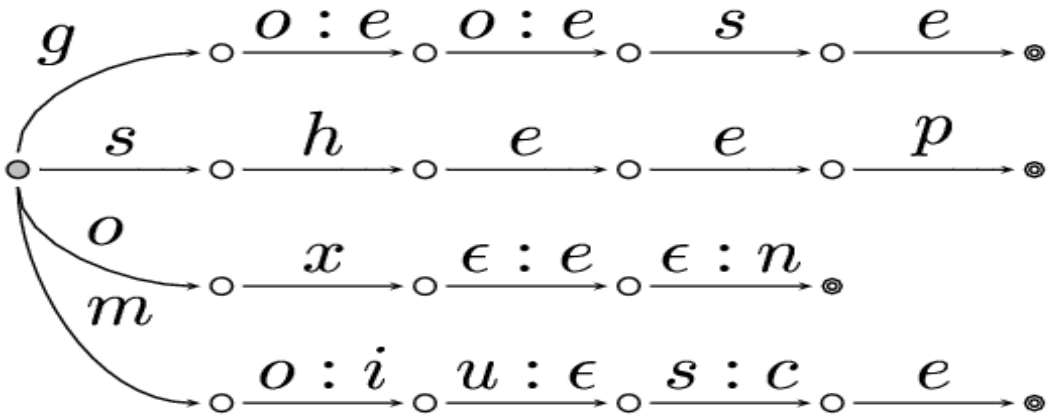
reg-noun	irreg-sg-noun	irreg-pl-noun
fox	goose	g o:e o:e s e
cat	sheep	sheep
dog	mouse	m o:i u:ɛ s:c e



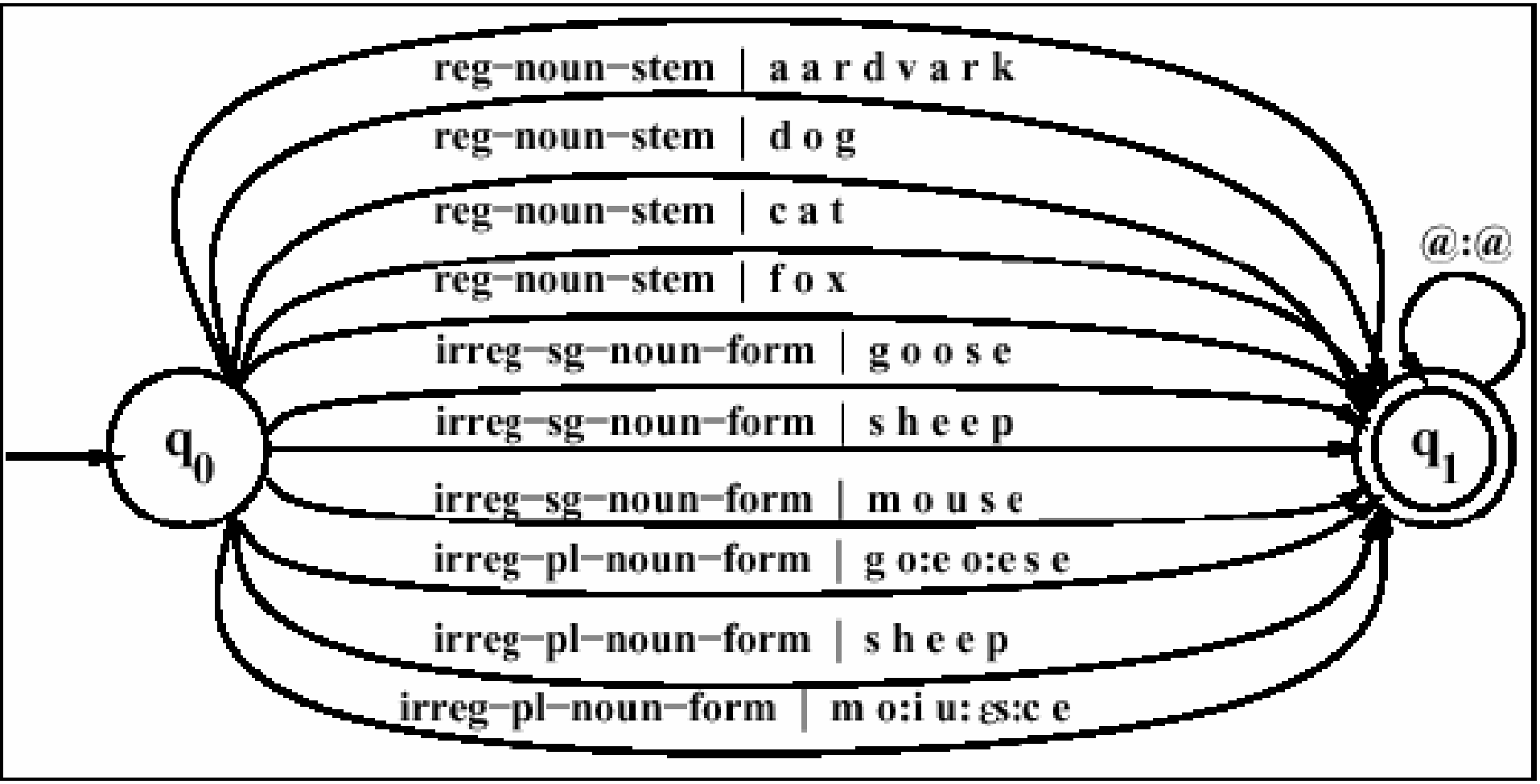
Shorthand notation:



Adding ϵ-moves:



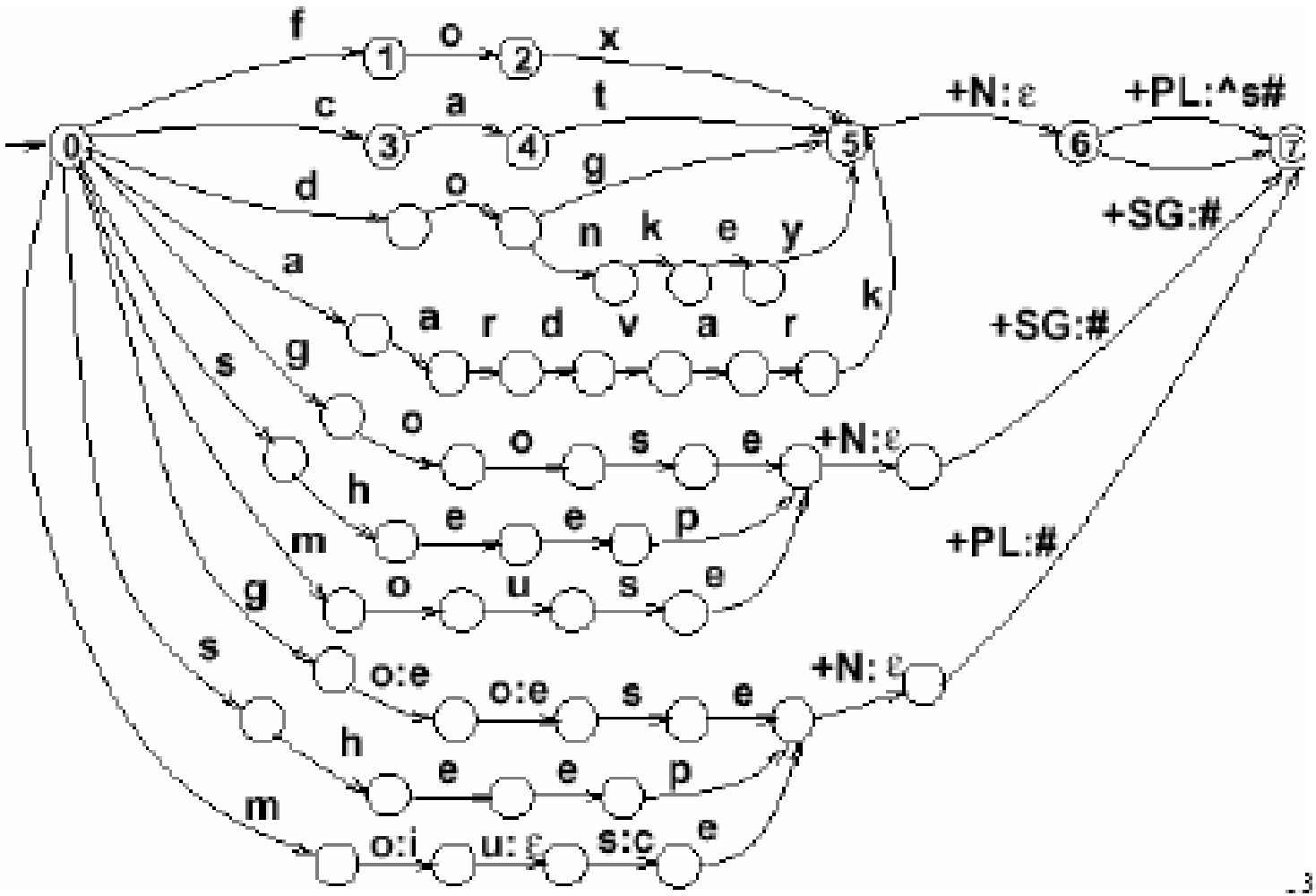
Το λεξικό του ΜΠΚ



Συνδυασμός των 2 ΜΠΚ

- Οι δύο ΜΠΚ (για το λεξικό και τους μορφολογικούς κανόνες) μπορούν να ενωθούν σειριακά (**cascaded**)
- Η έξοδος του ΜΠΚ για το λεξικό οδηγείται στην είσοδο του ΜΠΚ των μορφολογικών κανόνων
- Με την διαδικασία της **σύνθεσης** (composition) είναι δυνατό από πολλούς ΜΠΚ να δημιουργήσουμε έναν μοναδικό ΜΠΚ δυο ταινιών που να αντιστοιχεί από τη λεξικολογική μορφή στην επιφανειακή μορφή

Ο συνδυασμένος ΜΠΚ



Ορθογραφικοί κανόνες

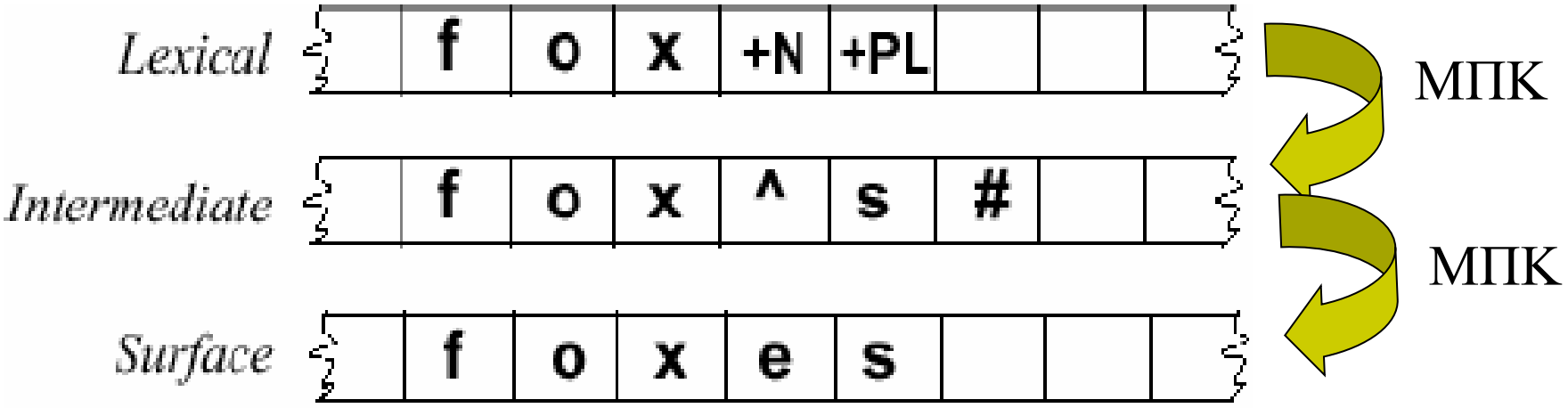
- Το προηγούμενο FST θα αποδεχόταν τη λέξη *foxs* και θα απέρριπτε τη λέξη *foxes*
- Χρειάζεται να χειριστούμε τις αλλαγές που συμβαίνουν συχνά στα όρια των μορφημάτων
- Αυτό γίνεται με τους ορθογραφικούς κανόνες
  - *Π.χ.* εισάγεται ένα *e* μετά τα *-s, -z, -x, -ch, -sh* και πριν από το *-s* *(watch/watches, ax/axes)*

    a -> b/c__d σημαίνει το *a* γίνεται *b* όταν βρίσκεται μεταξύ των *c* και *d*

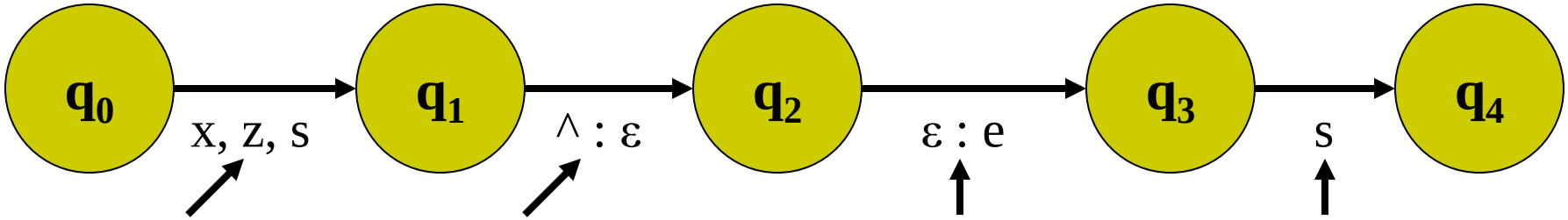
    ε ->e/{x,s,z}^__s#.
  - *Το -y* γίνεται *-ie* πριν το *-s* *(try/tries, sky/skies)*

Ορθογραφικοί κανόνες και ΜΠΚ

- Ένας ορθογραφικός κανόνας θεωρεί το εξωτερικό (δεύτερο) επίπεδο του προηγούμενου ΜΠΚ (το επίπεδο της συνένωσης των μορφημάτων) σαν ενδιάμεσο (intermediate) επίπεδο.
- Παράγει ένα καινούριο επιφανειακό επίπεδο, το οποίο αναπαριστά μια συνένωση καινούριων μορφημάτων που είναι ορθογραφικά σωστή.



Δημιουργία του ΜΠΚ για τον προηγούμενο ορθογραφικό κανόνα

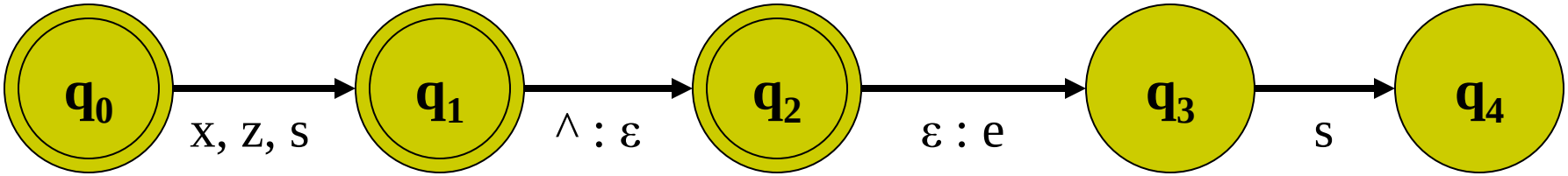


Μόλις μέσα στην λέξη μου φτάσω σε γράμμα x, z ή s

Και το γράμμα αυτό αποτελεί τέλος μορφήματος, πρέπει να σβηστεί το όριο του μορφήματος

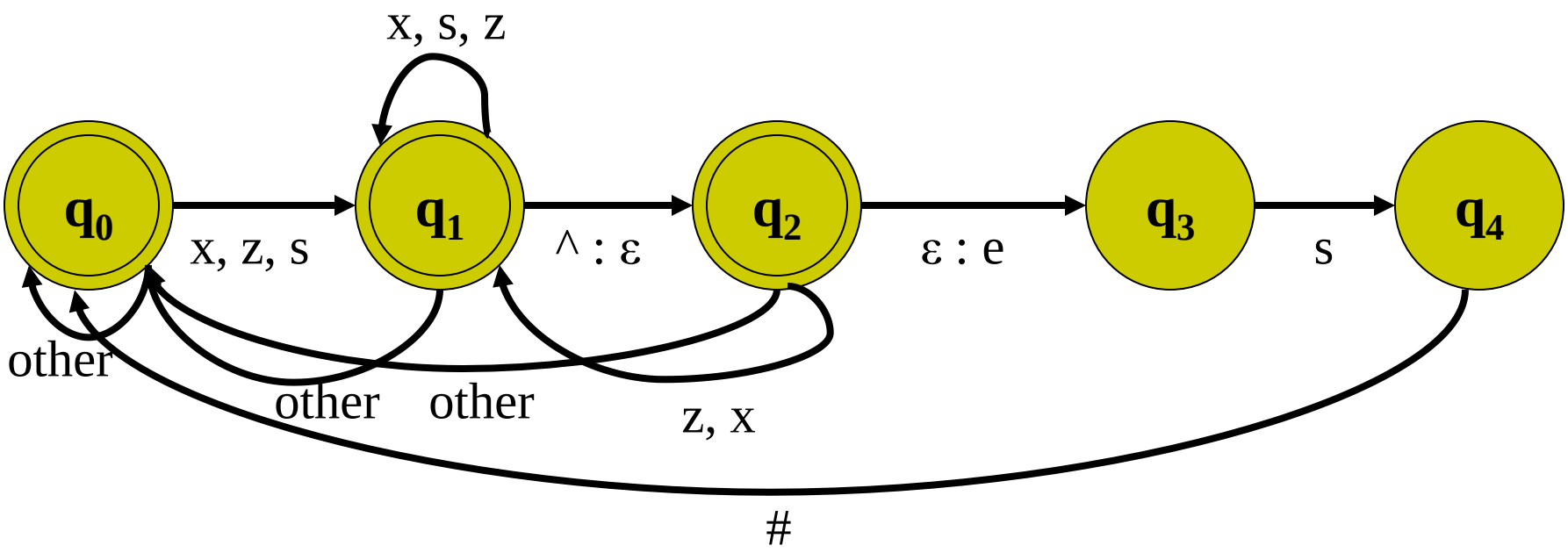
Προστίθεται το γράμμα e

Ακολουθεί το γράμμα s



Η **q0** είναι τελική κατάσταση για όλα τα default pairs που είναι άσχετα με τον κανόνα. Η **q1** είναι τελική για τις λέξεις που περιλαμβάνουν x, z, s. Η **q2** είναι τελική κατάσταση για όσες λέξεις το x, z, ή s είναι τέλος μορφήματος.

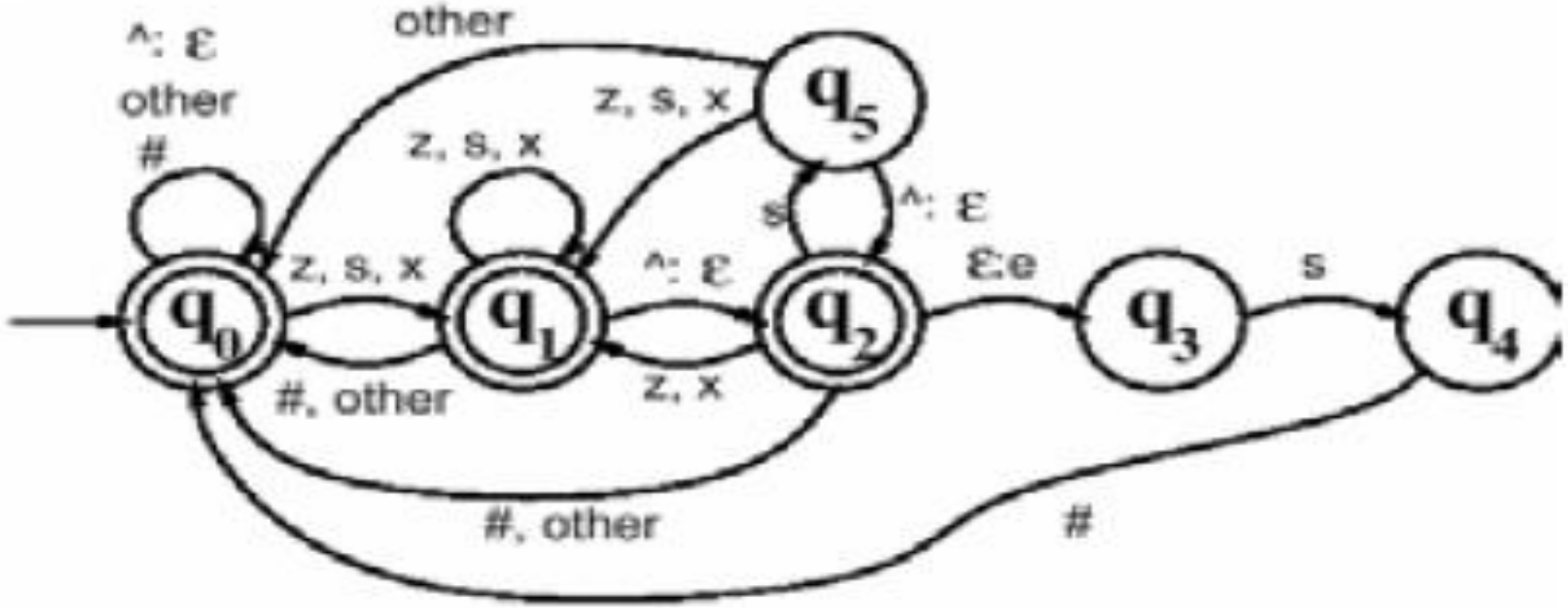
Δημιουργία του ΜΠΚ για τον προηγούμενο ορθογραφικό κανόνα (συν)



Πρέπει να εξασφαλιστεί η κάλυψη και των υπόλοιπων (άσχετων με τον συγκεκριμένο κανόνα) περιπτώσεων.

Ορθογραφικοί κανόνες και ΜΠΚ

Πρέπει να προστεθεί μια ακόμα κατάσταση (η q_5), η οποία θα εξασφαλίζει ότι σε κάθε περίπτωση που θα ικανοποιείται το πριν και μετά περιβάλλον, πάντα θα προστίθεται το e. Έτσι από την q_2, (δηλ. πριν προστεθεί το e) αν έρθει s πάω στην q_5, και από εκεί δεν μπορεί να έχω τέλος λέξης (#).



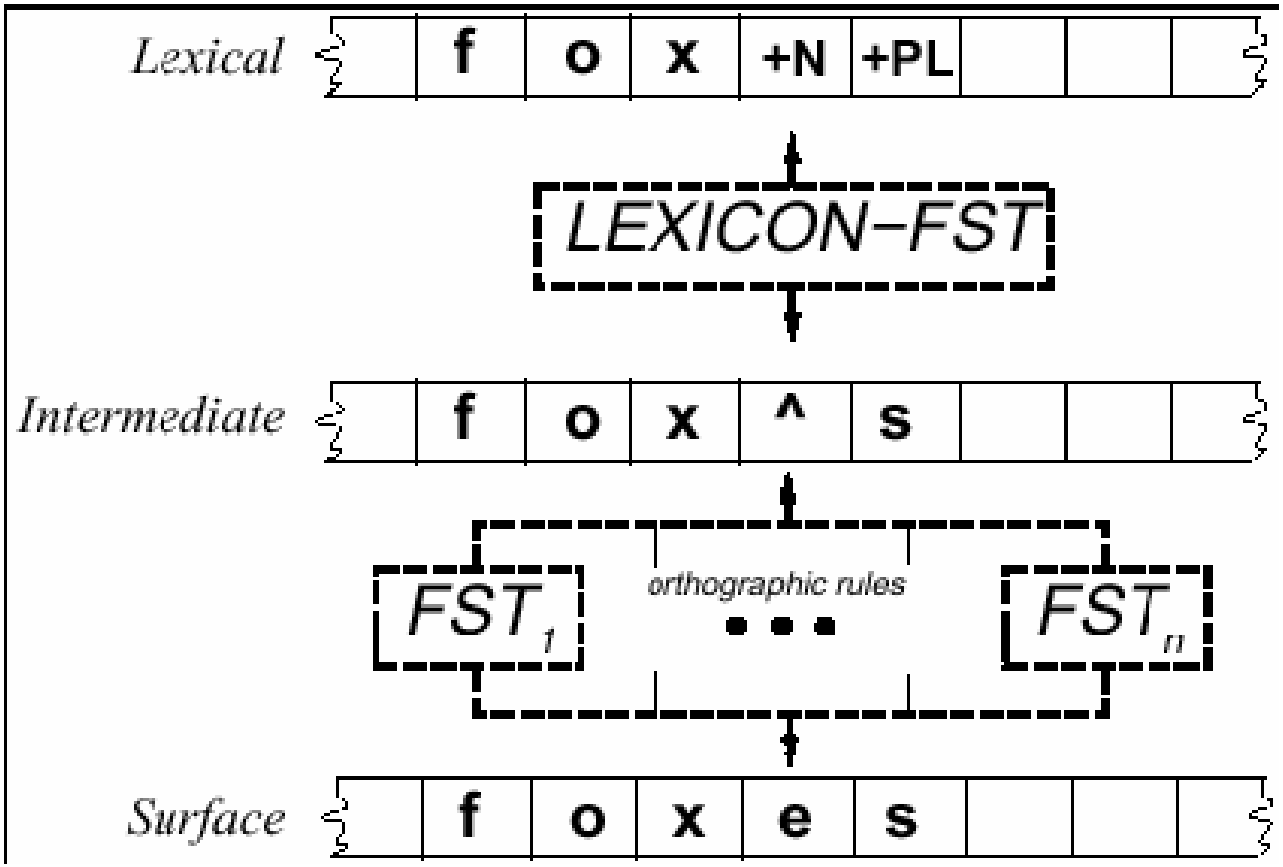
Πίνακας Μετάβασης

State/ Input	s:s	x:x	z:z	^:ε	ε:e	#	other
q0:	1	1	1	0	-	0	0
q1:	1	1	1	2	-	0	0
q2:	5	1	1	0	3	0	0
q3	4	-	-	-	-	-	-
q4	-	-	-	-	-	0	-
q5	1	1	1	2	-	-	0

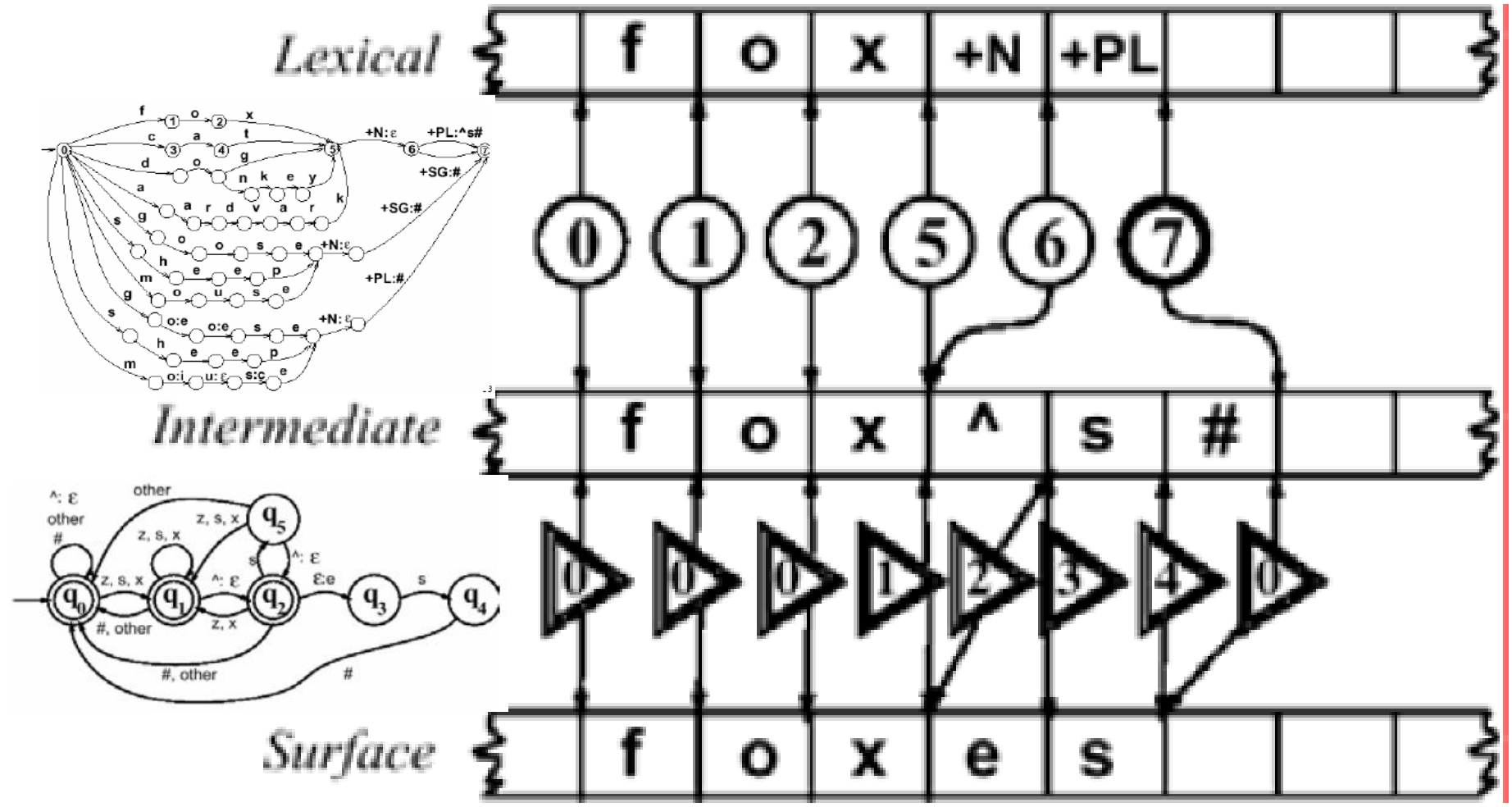
Συνδυασμός των ΜΠΚ του Λεξικού και των Κανόνων

- Αρχικά ο ΜΠΚ λεξικού αντιστοιχίζει το λεξιλογικό επίπεδο στο ενδιάμεσο επίπεδο (απλή συνένωση μορφημάτων)
- Ένας αριθμός ΜΠΚ κανόνων τρέχουν παράλληλα (ή ως cascade) και αντιστοιχίζουν το ενδιάμεσο επίπεδο στο επιφανειακό επίπεδο
- Ο ΜΠΚ λεξικού και οι ΜΠΚ των ορθογραφικών κανόνων σχηματίζουν ένα cascade.
  - top-down (generation)
  - bottom-up (parsing)

Συνδυασμός των ΜΠΚ του Λεξικού και των Κανόνων



Συνδυασμός των ΜΠΚ του Λεξικού και των Κανόνων



Ανάλυση-Παραγωγή

- Η ανάλυση (parsing) είναι πιο περίπλοκη από την παραγωγή (generation) λόγω της ασάφειας.
  - Π.χ. η λέξη *foxes* μπορεί να αναλυθεί
    ως *fox+V+3SG* *(foxes: κοροϊδεύει)*
    και *fox+N+PL* *(foxes: αλεπούδες)*
- Η αποσαφήνιση (disambiguation) δεν μπορεί να επιλυθεί χωρίς τα συμφραζόμενα
  - Ο ΜΠΚ πρέπει να δημιουργήσει και τις δύο αναλύσεις
- Επίσης, προκαλούνται ασάφειες εξαιτίας του ε ή λόγω πολλαπλών πιθανών μονοπατιών

Two-Level Morphology Software

- PC-KIMMO
  - http://www.sil.org/pckimmo/
- Μορφολογικές περιγραφές για τα Αγγλικά και άλλες γλώσσες είναι διαθέσιμες

Stemming

- Stemming ονομάζεται η εύρεση της ρίζας μιας λέξης
  - marsupials -> marsupial
  - trying -> try
- Σε πολλές εφαρμογές (ιδίως ανάκτησης πληροφορίας) δεν είναι απαραίτητη πλήρης μορφολογική ανάλυση (που απαιτεί πολύπλοκους κανόνες και λεξικά)
- Απλά αρκεί να βρεθεί ότι δύο λέξεις έχουν την ίδια ρίζα.

Porter Stemmer (1980)

- http://www.tartarus.org/~martin/PorterStemmer/
- Πρόκειται για ευρέως διαδεδομένο stemmer
- Βασίζεται σε έναν αριθμό απλών κανόνων
  - ATIONAL -> ATE (relational -> relate)
  - ING-> ε (knowing -> know)
- Οι stemmers δεν έχουν απόλυτη ακρίβεια
  - Λάθη του Porter stemmer:
    - organization -> organ
    - policy -> police
  - Περιπτώσεις που ο Porter stemmer δεν καλύπτει:
    - matrices -> matrix
    - explain -> explanation

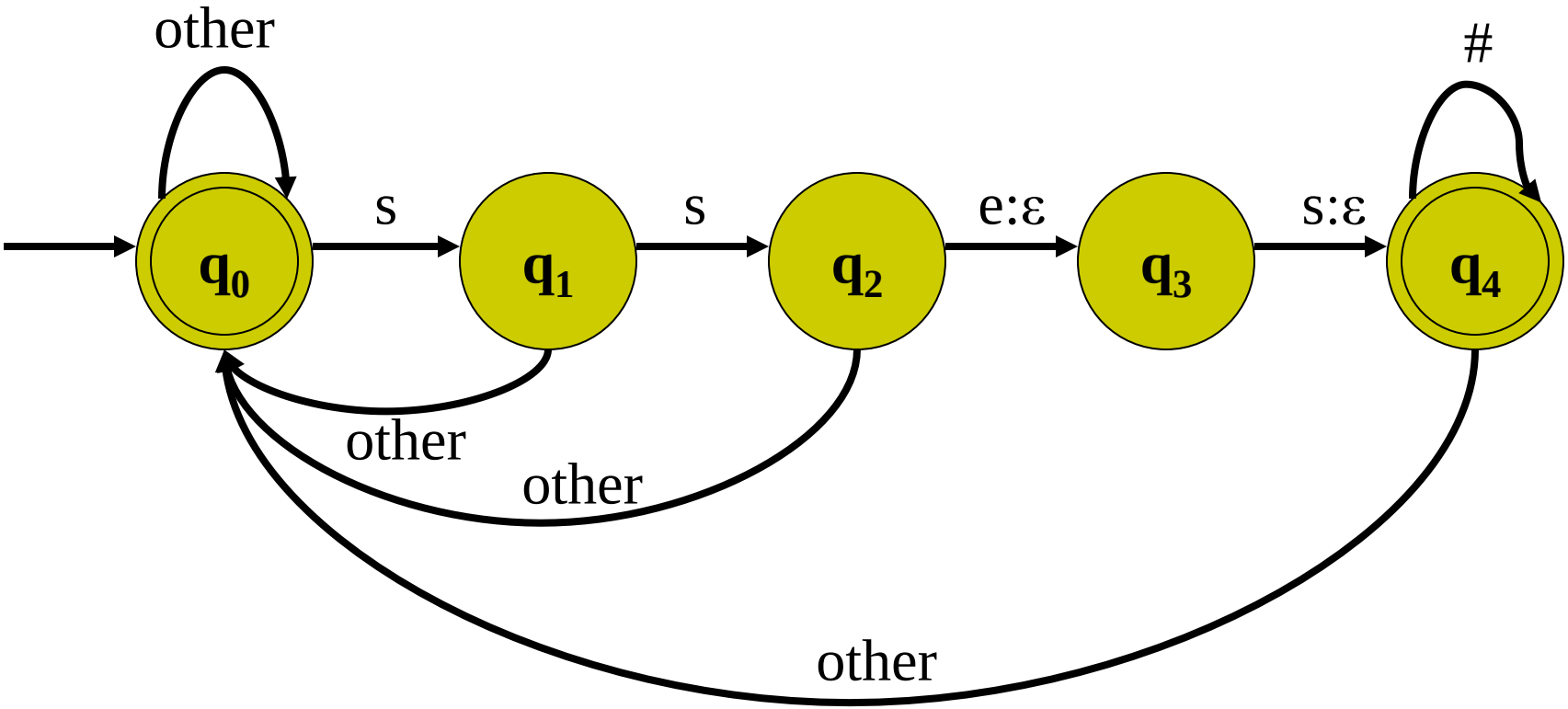
Άσκηση

- Στον Porter Stemmer εφαρμόζονται οι ακόλουθοι κανόνες για την λημματοποίηση μιας λέξης:
  - Xsses -> Xss
    - caresses -> caress
  - Xies -> Xi
    - ponies -> poni
  - Xss -> Xss
    - caress -> caress
  - Xs -> X
    - cats -> cat
  - Xeed -> Xee , (|X|>1)
    - feed -> fee, agreed -> agree
  - Xed -> X , (X contains vowel)
    - plastered -> plaster, bled -> bled
  - Xing -> X , (X contains vowel)
    - motoring -> motor, sing -> sing

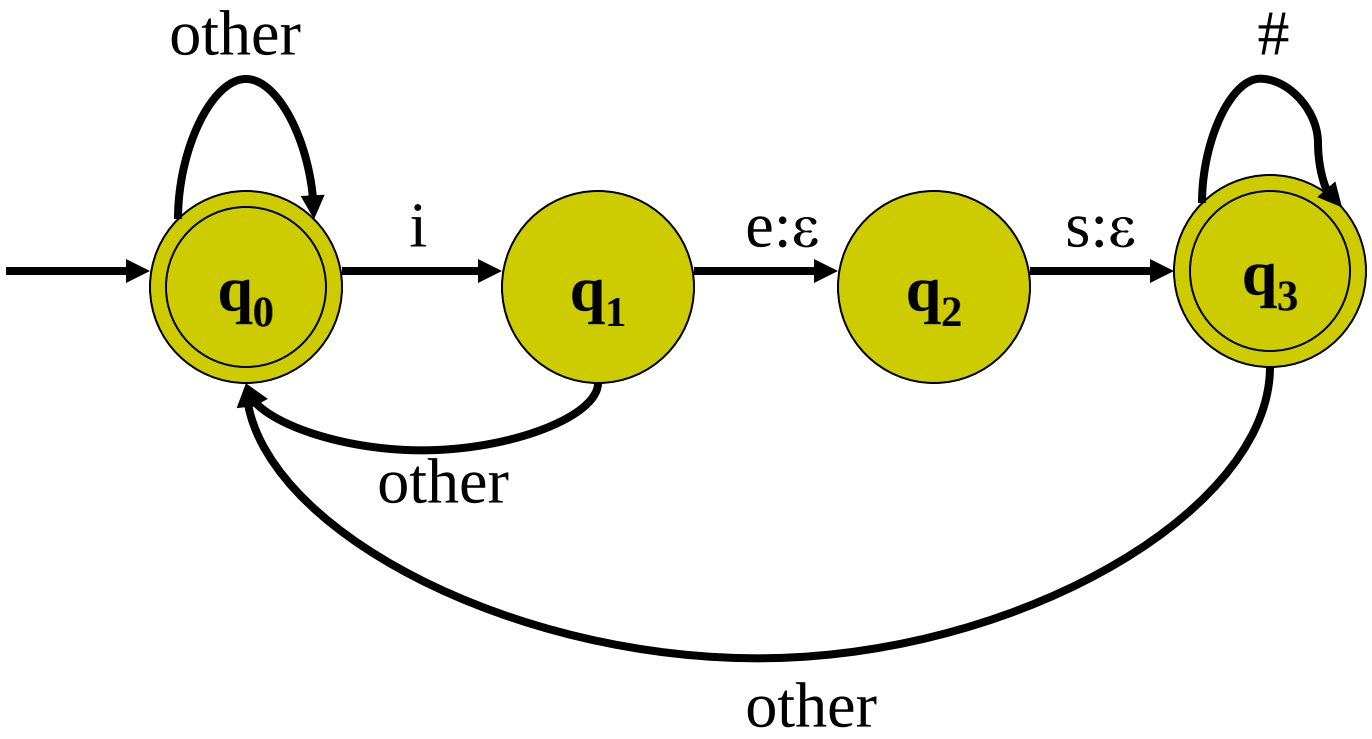
X: οποιαδήποτε ακολουθία χαρ/ρων

Να υλοποιηθούν οι κανόνες αυτοί με ΜΠΚ.

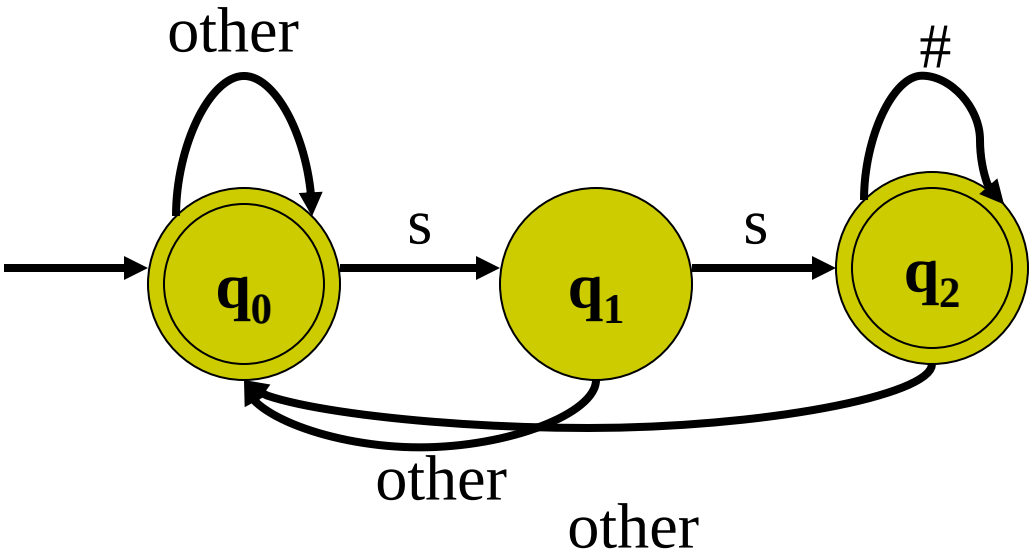
1ος κανόνας: Xsses -> Xss



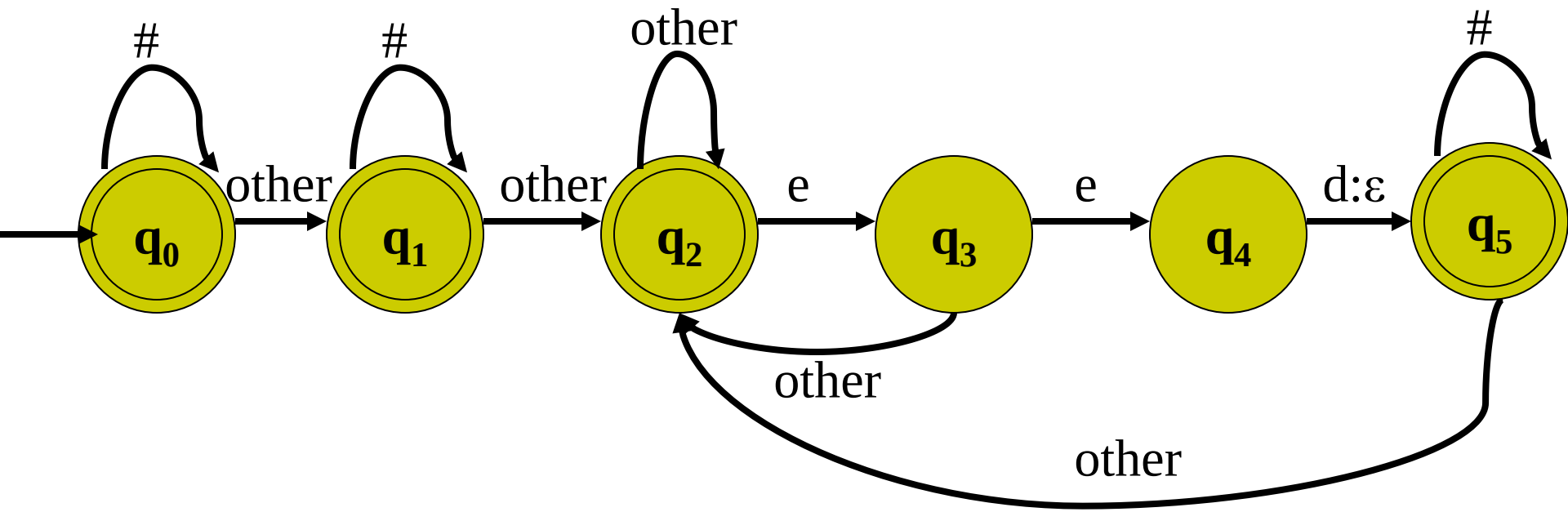
2ος κανόνας: Xies -> Xi



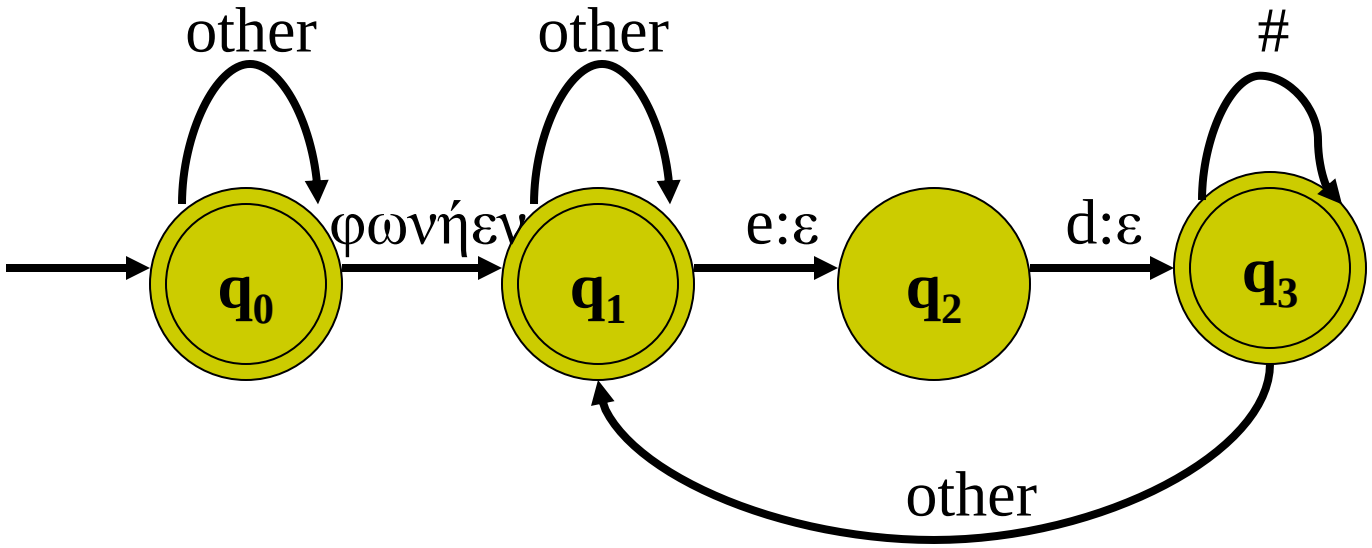
3ος κανόνας: Xss -> Xss



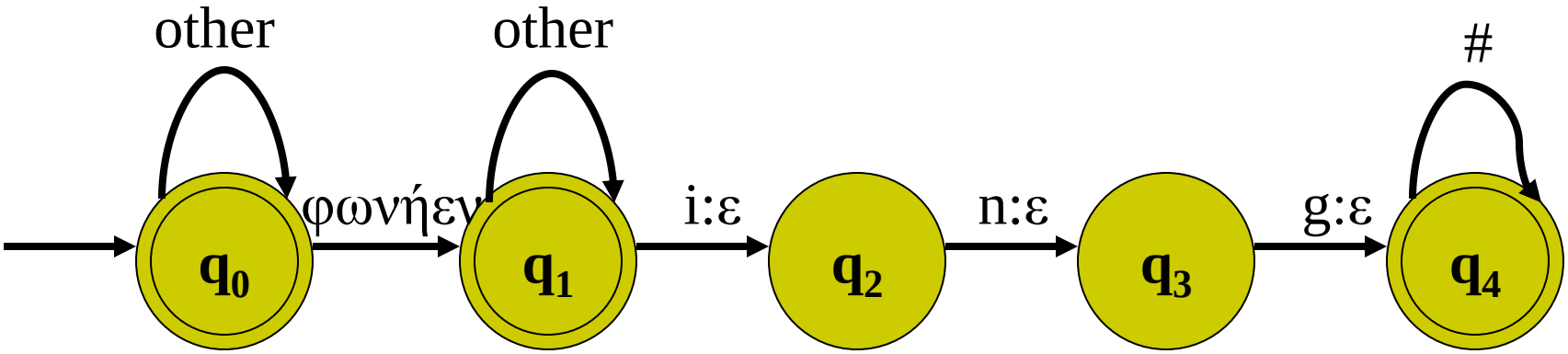
4ος κανόνας: Xeed -> Xee, |X|>1



5ος κανόνας: Xed -> X, (το X περιέχει φωνήεν)



6ος κανόνας: Xing -> X, (το X περιέχει φωνήεν)



Ο αλγόριθμος Soundex

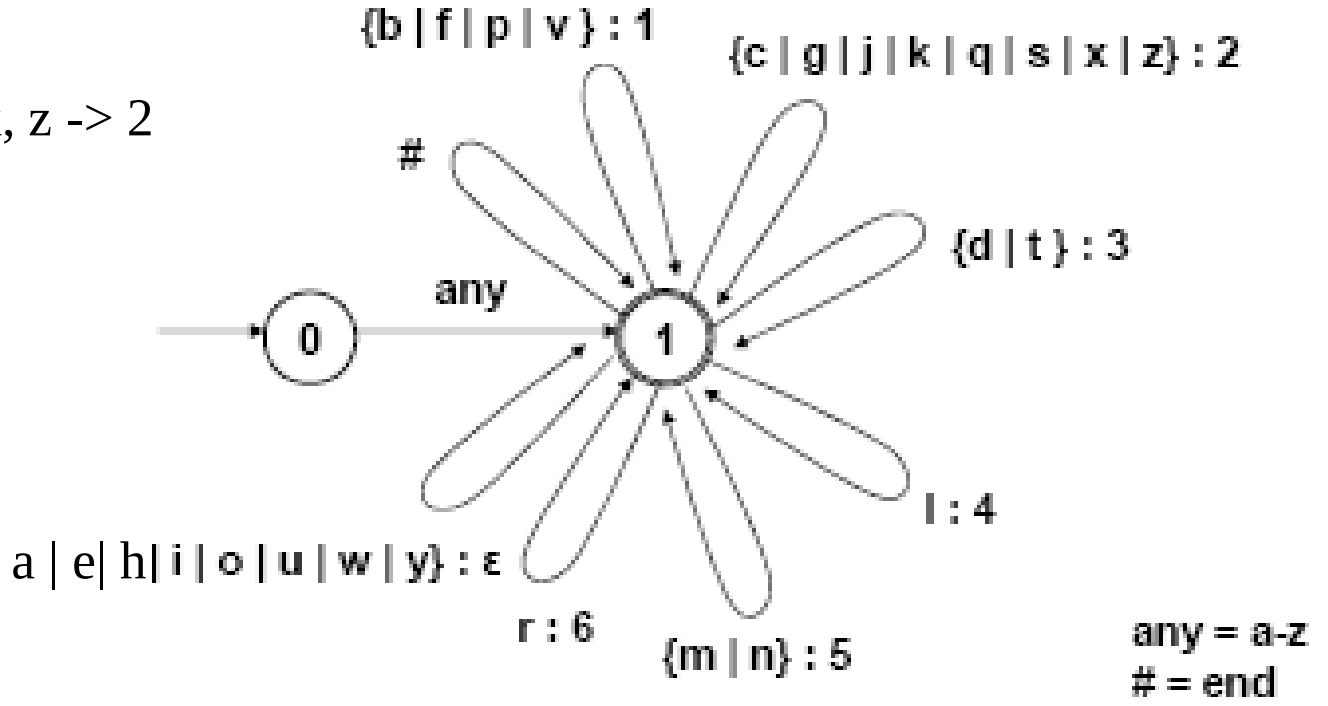
- Ο αλγόριθμος Soundex μετατρέπει κύρια ονόματα σε κωδικούς:
  - Jurafsky -> J612
  - Jarofski -> J612
  - Clinton -> C453
  - Bush -> B200

Ο αλγόριθμος

- Κράτησε το πρώτο γράμμα του ονόματος και αγνόησε όλες τις εμφανίσεις των μη-αρχικών a, e, h, i, o, u, w, y
- Αντικατέστησε τα υπόλοιπα γράμματα με τους ακόλουθους αριθμούς:
  - b, f, p, v -> 1
  - c,g, j, k, q, s, x, z -> 2
  - d,t -> 3
  - l -> 4
  - m, n -> 5
  - r -> 6
- Αντικατέστησε τυχόν ακολουθίες ίδιων αριθμών με ένα απλό αριθμό (666->6)
- Μετέτρεψε στη μορφή ‘Γράμμα Ψηφίο Ψηφίο Ψηφίο’ αγνοώντας τα ψηφία μετά το τρίτο (αν χρειάζεται) ή προσθέτοντας μηδενικά στο τέλος (αν χρειάζεται)

Υλοποίηση ΜΠΚ (1ο βήμα)

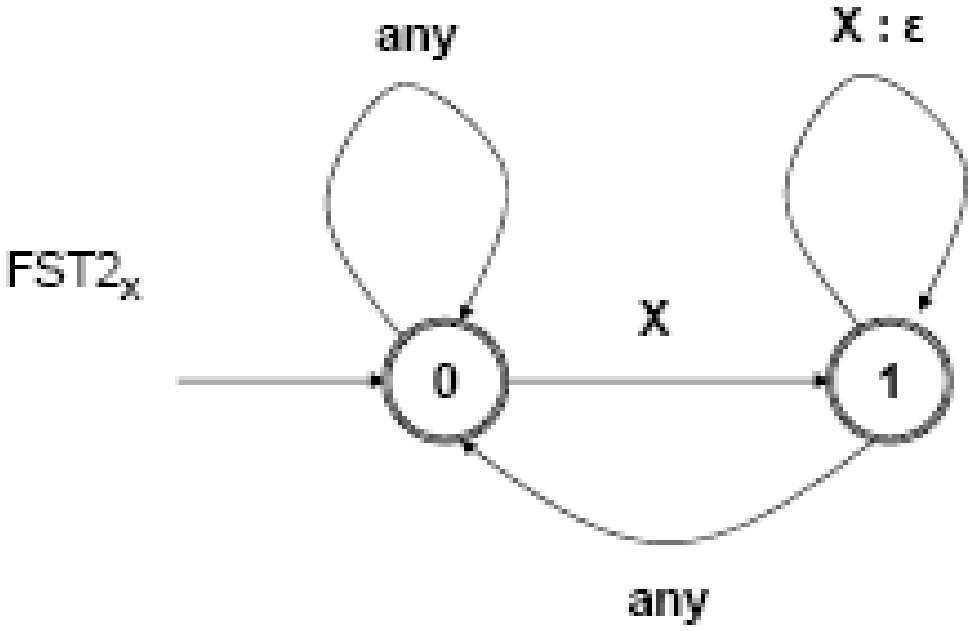
- Κράτησε το πρώτο γράμμα του ονόματος και αγνόησε όλες τις εμφανίσεις των μη-αρχικών a, e, h, i, o, u, w, y
- Αντικατέστησε τα υπόλοιπα γράμματα με τους ακόλουθους αριθμούς:
  - b, f, p, v -> 1
  - c,g, j, k, q, s, x, z -> 2
  - d,t -> 3
  - l -> 4
  - m, n -> 5
  - r -> 6



FST1

Υλοποίηση ΜΠΚ (2ο βήμα)

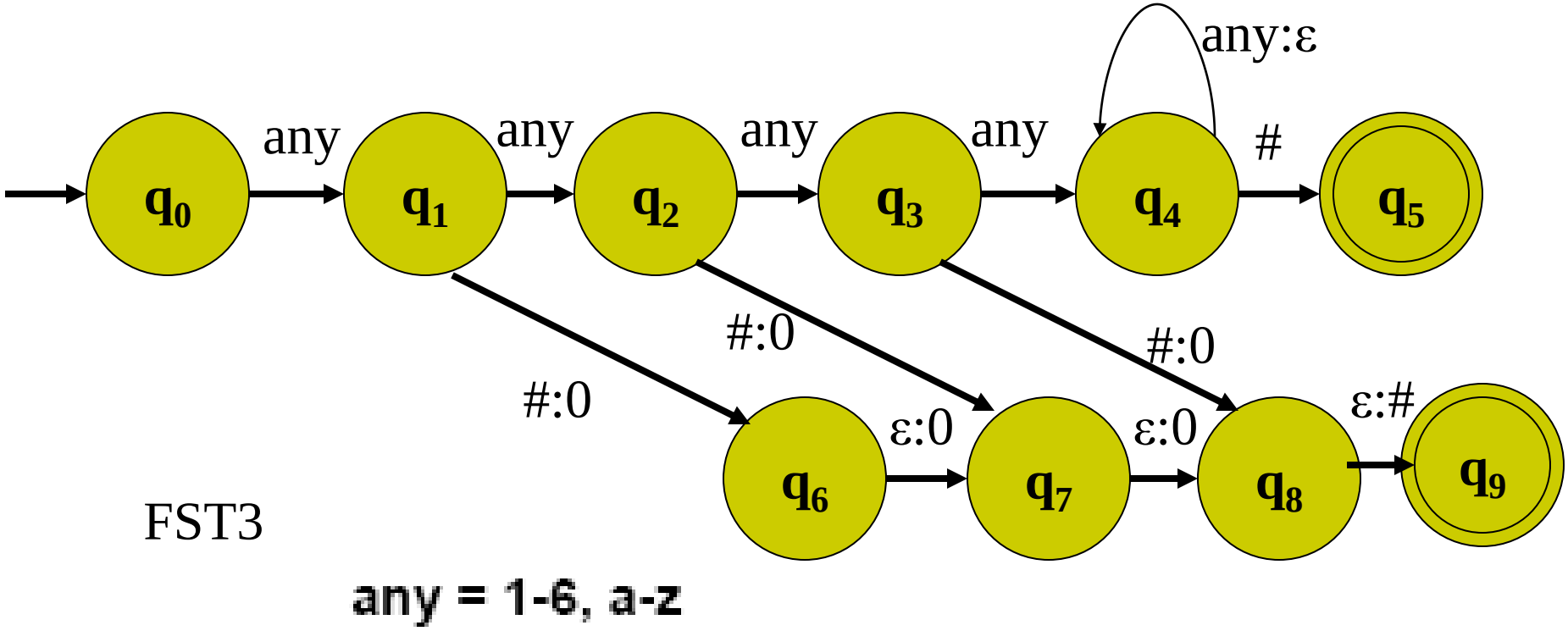
- Αντικατέστησε τυχόν ακολουθίες ίδιων αριθμών με έναν απλό αριθμό (666->6)



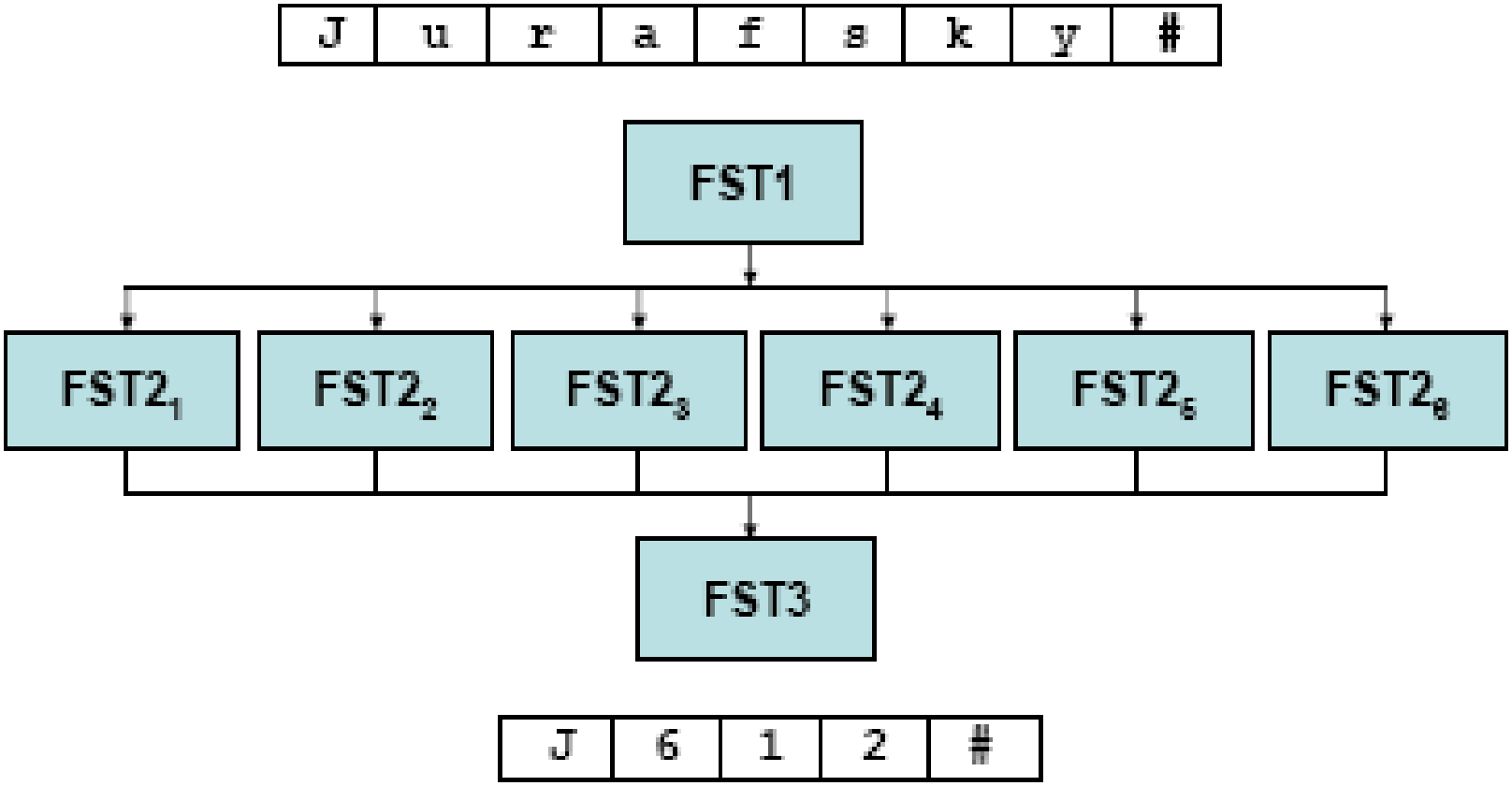
X = 1-6
any = a-z, #, 1-6 (-X)

Υλοποίηση ΜΠΚ (3ο βήμα)

- Μετέτρεψε στη μορφή ‘Γράμμα Ψηφίο Ψηφίο Ψηφίο’ αγνοώντας τα ψηφία μετά το τρίτο (αν χρειάζεται) ή προσθέτοντας μηδενικά στο τέλος (αν χρειάζεται)



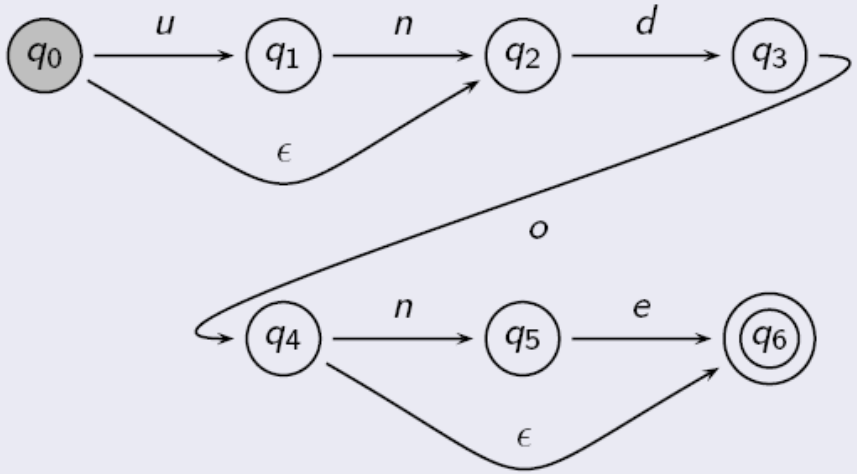
Σχεδιασμός


J u r a f s k y #
FST1
FST2$_1$
FST2$_2$
FST2$_3$
FST2$_4$
FST2$_5$
FST2$_6$
FST3
J 6 1 2 #

Άσκηση για ΑΠΚ

Example: Automata with ϵ-moves

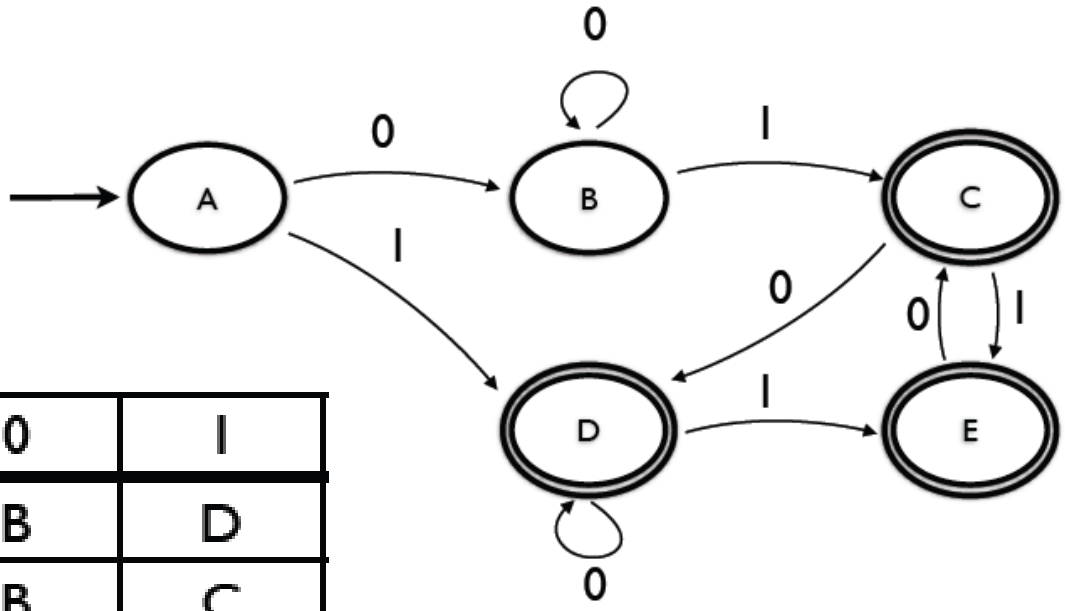
The language accepted by the following automaton is {*do*, *undo*, *done*, *undone*}:



Ασκηση
Σχεδιάστε το παρακάτω ΑΠΚ

δ_3	0	1
A	B	D
B	B	C
C	D	E
D	D	E
E	C	

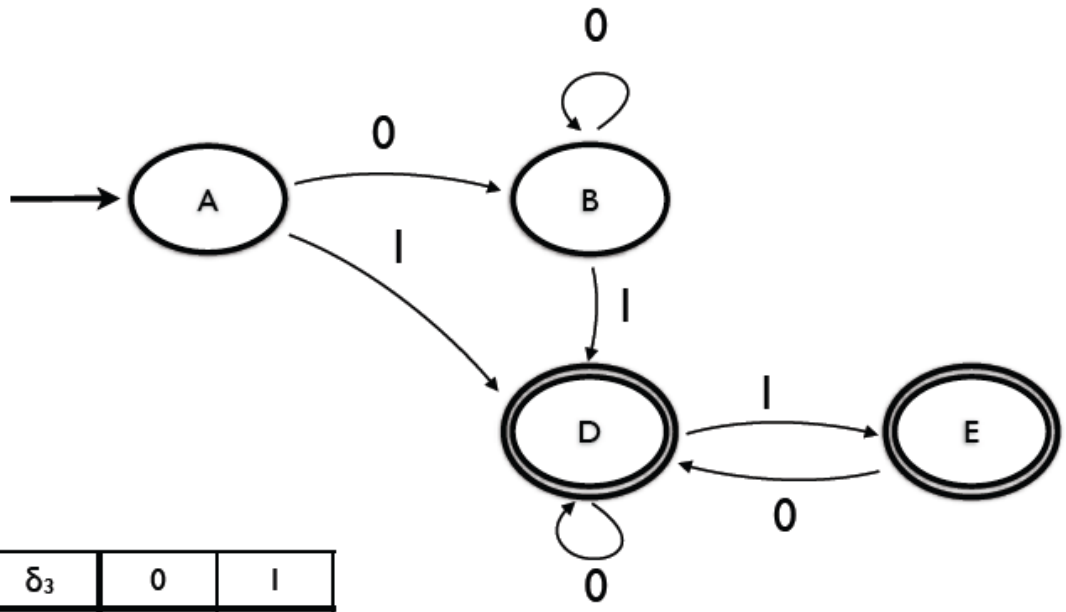
Λύση (1/3)



δ_3	0	1
A	B	D
B	B	C
C	D	E
D	D	E
E	C	

Οι καταστάσεις C και D είναι ισοδύναμες

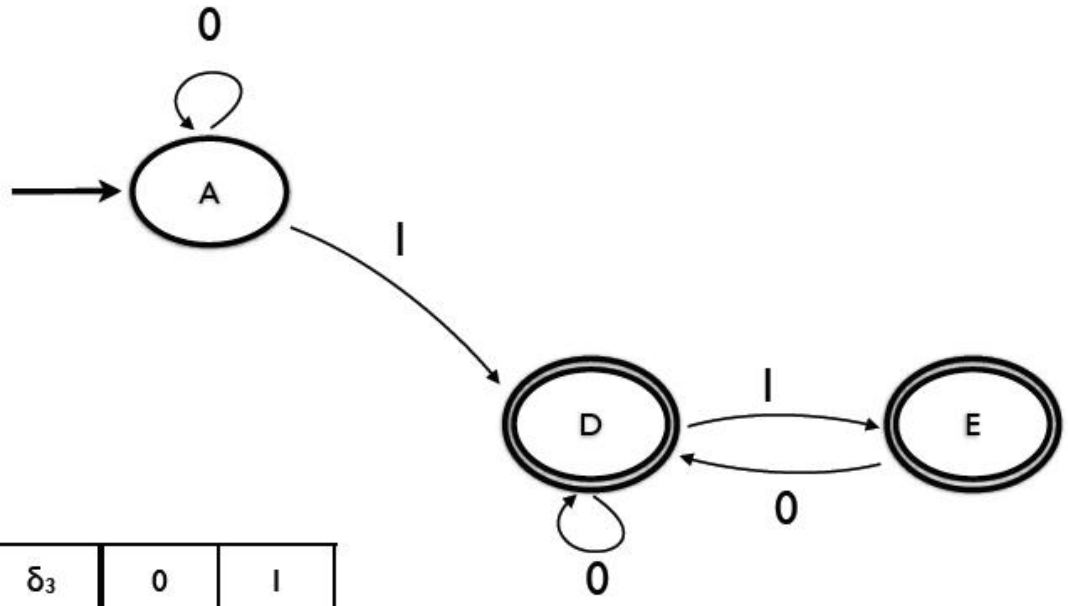
Λύση (2/3)



δ_3	0	1
A	B	D
B	B	D
D	D	E
E	D	

We can thus remove C and redirect all its incoming edges to D

Λύση (3/3)



δ_3	0	1
A	A	D
D	D	E
E	D	

And we‘re done!