«Τεχνογλωσσία VIII»
Εξαγωγή πληροφοριών από κείμενα

Σεμινάριο 4: Συντακτική Ανάλυση

Ευάγγελος Καρκαλέτσης, Γεώργιος Πετάσης

Εργαστήριο Τεχνολογίας Γνώσεων & Λογισμικού,

Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών, Ε.Κ.Ε.Φ.Ε. "Δημόκριτος"

Τηλ.: 210-6503197, Fax: 210-6532175, {vangelis, petasis}@iit.demokritos.gr

Ακαδημαϊκό Έτος: 2013 – 2014

Οι διαφάνειες αυτού του μαθήματος βασίζονται
στα κεφάλαια 12 και 13 του βιβλίου:

«Speech and Language Processing»
των D. Jurafsky και J.H. Martin, 2η έκδοση, Pearson, 2009

Το βιβλίο δεν απαιτείται για το μάθημα αυτό.

Μερικά παραδείγματα βασίζονται σε διαφάνειες
του Δρ. Ίων Ανδρουτσόπουλου, ΟΠΑ.
http://www.aueb.gr/users/ion/

Σύνταξη

- Ο τομέας της γλωσσολογίας που μελετά τη δομή των προτάσεων
  - Δηλαδή ποιές σχέσεις συνδέουν μια ακολουθία
  - Σε καμία φυσική γλώσσα οι προτάσεις δεν αποτελούν τυχαία παράθεση λέξεων ή ομάδων λέξεων
  - Ύπαρξη κανόνων → δόμηση πρότασης
- Συντακτικοί κανόνες: καθολικοί ή όχι

  π.χ. α) Π → ΟΦ + ΡΦ

  β) γράφω/ εγώ γράφω, ενώ I write/*write

Γλωσσική ικανότητα και πλήρωση (1)

- Γλωσσική ικανότητα: γενικά η γνώση του φυσικού ομιλητή για τη γλώσσα του, μέρος της οποίας είναι και η γραμματική
  - Ικανότητα → γραμματικότητα πρότασης, γνώση δομής συστήματος
- Γλωσσική πλήρωση: η γλωσσική συμπεριφορά του ομιλητή κατά την επικοινωνία
    - Η Μαρία ξεκίνησε να διαβάζει το βιβλίο.
    - *Κώστας ο βάζο έσπασε το.
    - *Ο σκύλος τραγούδησε τα κίτρινα δάπεδα.
    - *Κοντεύω να φτάσεις.

Γλωσσική ικανότητα και πλήρωση (2)

- Η διάκριση γραμματικών από μη-γραμματικές προτάσεις είναι μέρος της γλωσσικής μας ικανότητας. Αυτό αποτελεί το γλωσσικό μας αίσθημα ή γλωσσική διαίσθηση

- Η γλωσσική πλήρωση, από την άλλη μεριά, επιτρέπει τόσο γραμματικές όσο και μη-γραμματικές προτάσεις και δεν κάνει αυτή τη διάκριση

Συντακτική ανάλυση (1)

- Η μετατροπή μιας πρότασης φυσική γλώσσας σε μια ιεραρχική δομή
  - Η οποία ανταποκρίνεται στην διασύνδεση των δομικών στοιχείων της πρότασης
- Η ανάλυση μπορεί να επιστρέψει περισσότερες από μία δομές (parses)
- Η πιο απλή μορφή δομής είναι ένα συντακτικό δέντρο (syntax tree)

Συντακτική ανάλυση (2)

- Συνήθως τα υπάρχοντα συστήματα έχουν δύο συστατικά:
  - **Γραμματική** (grammar): ρητή αναπαράσταση των συντακτικών κανόνων της γλώσσας
    - Δηλωτικοί φορμαλισμοί που ορίζουν τις έγκυρες προτάσεις μιας γλώσσας, αλλά δεν καθορίζουν πως θα γίνει η αναγνώριση και η παραγωγή συντακτικών δομών
  - **Αναλυτής** (parser): αναλύει τις προτάσεις εισόδου, συγκρίνοντάς τες με την γραμματική, και παράγει συντακτικές δομές

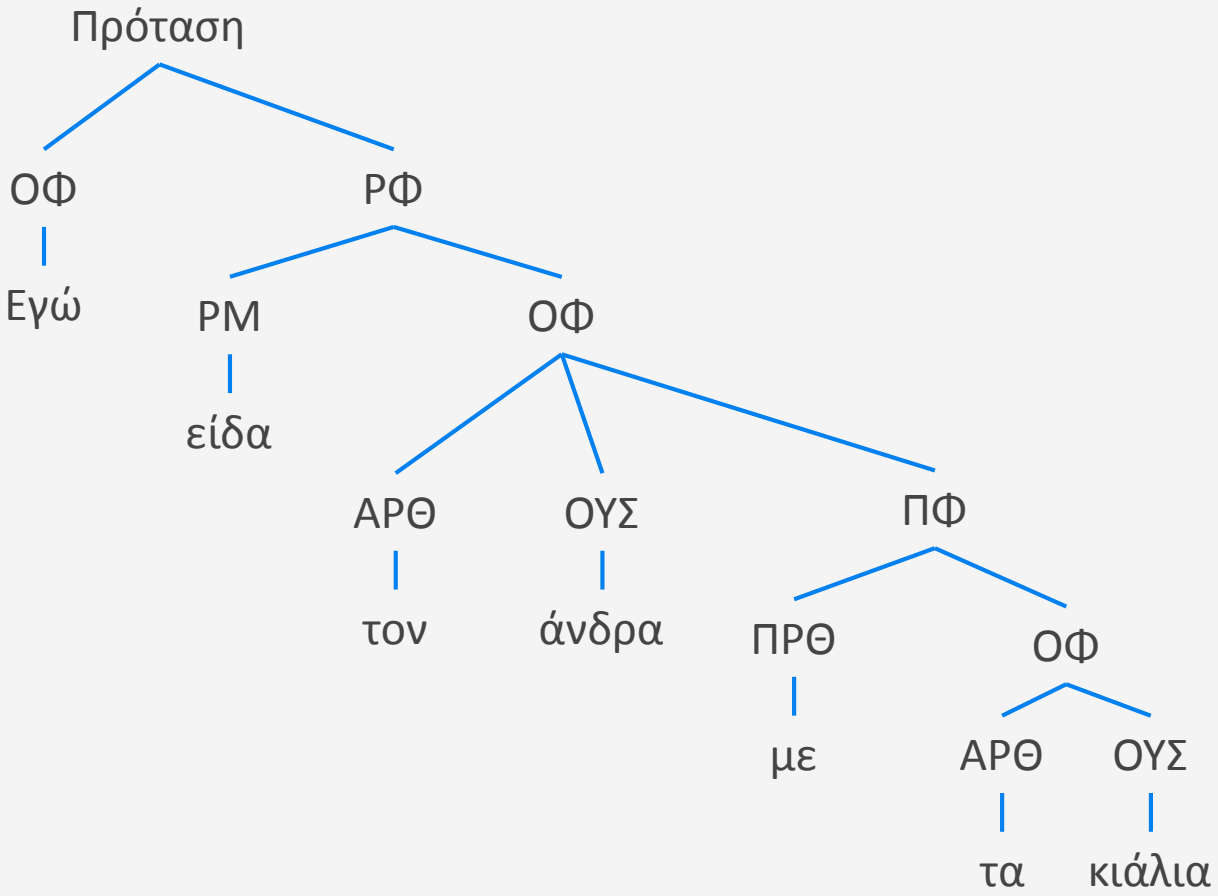
Τι χρειαζόμαστε για την ανάλυση; (1)

- Τι γλωσσική πληροφορία χρειαζόμαστε για την συντακτική ανάλυση;
  - Λέξεις
    - Κατηγορίες: σύνολα λέξεων που συμπεριφέρονται όμοια
    - Μέρη του λόγου: Ουσιαστικά, ρήματα, επίθετα, προθέσεις, κλπ.
  - Συστατικά (constituents):
    - Ομαδοποίηση λέξεων σε μεγαλύτερες ενότητες, οι οποίες συμπεριφέρονται όμοια
      - Και έχουν ένα συγκεκριμένο μέρος του λόγου σαν «κύριο» (head)
    - Φράσεις: Ονοματική φράση με «κύριο» το ουσιαστικό, ρηματική φράση με «κύριο» το ρήμα, κλπ.

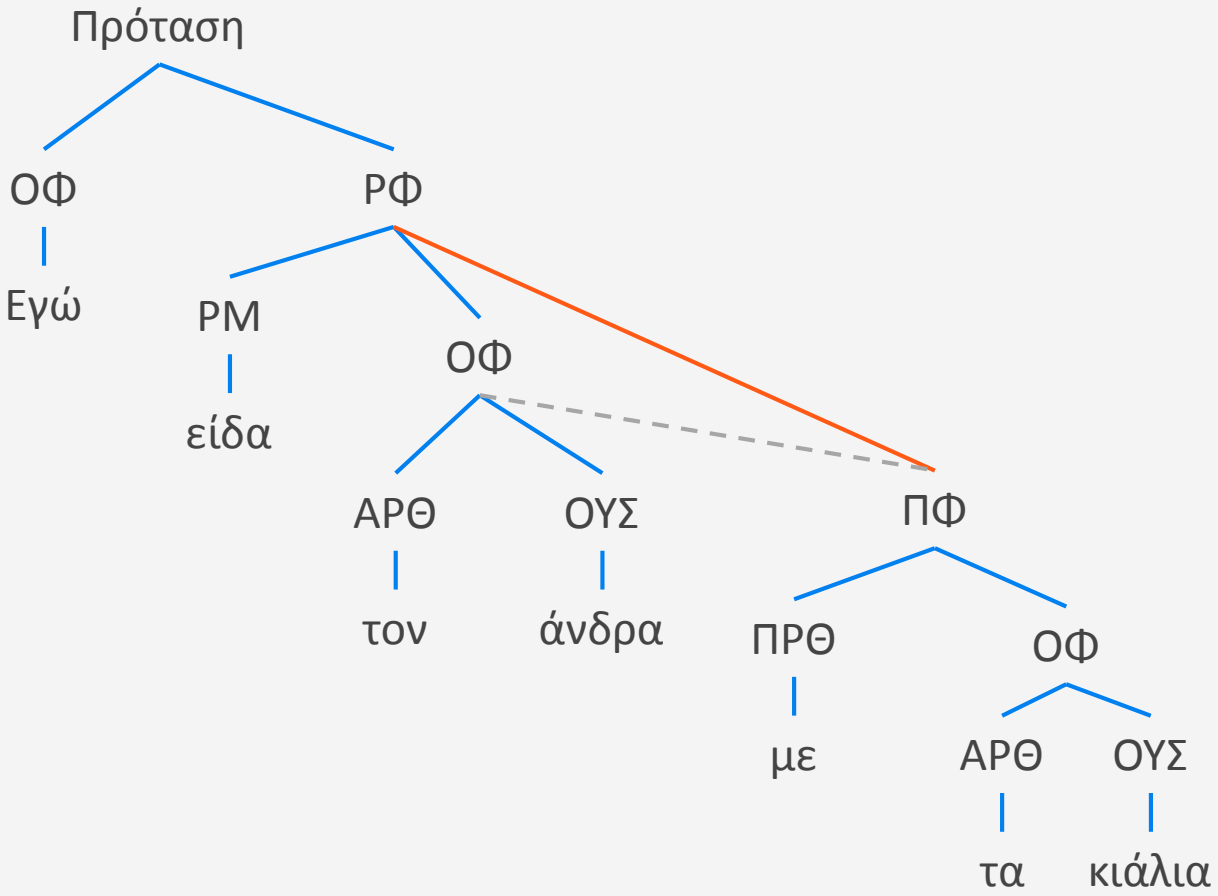
Τι χρειαζόμαστε για την ανάλυση; (2)

- Έχοντας:
  - Μορφολογική ανάλυση: ανάλυση λέξεων σε μορφήματα και προσφύματα
    - Με κανόνες, FSA, FST
  - Αναγνώριση μερών του λόγου
- Περιμένουμε από την συντακτική ανάλυση:
  - Να προσδιορίσει τα συστατικά, και πως σχετίζονται
  - Να προσδιορίσει αν μια πρόταση είναι γραμματικά σωστή
  - Να παραγάγει συντακτικές δομές

Παράδειγμα: Ανάλυση 1



Παράδειγμα: Ανάλυση 2



Γραμματικές ΦΓ (1)

NP → Det Nominal

NP → ProperNoun

Nominal → Noun | Nominal Noun

Διάζευξη – Ουσιαστικά δύο κανόνες

Det → a

Det → the

Noun → flight

Λεξικό – Στην πράξη πληροφορίες από την μορφολογική ανάλυση

- Τερματικά σύμβολα, Μη τερματικά σύμβολα
- Κανόνες $\alpha \rightarrow \beta$: ορίζουν τις δυνατές «παραγωγές»
- **Αρχικό σύμβολο**: ένα από τα μη τερματικά

Γραμματικές ΦΓ (2)

- Οι γραμματικές χρησιμοποιούνται:
  - Για την ανάλυση φυσικής γλώσσας
  - Για την παραγωγή φυσικής γλώσσας

- **Γλώσσα** της γραμματικής: οι ακολουθίες τερματικών συμβόλων που παράγονται από το αρχικό σύμβολο

Ιεραρχία γραμματικών του Chomsky (1)

- Τύπος 3: κανονικές γραμματικές (regular grammars)
  - Μορφή κανόνων
    - $A \rightarrow x$ και $A \rightarrow xB$ (δεξιά γραμμικές)
    - $A \rightarrow x$ και $A \rightarrow Bx$ (αριστερά γραμμικές)
    - x: (πιθανώς κενή) ακολουθία τερματικών συμβόλων
    - A, B: μεμονωμένα μη τερματικά σύμβολα

Ιεραρχία γραμματικών του Chomsky (2)

- Τύπος 2: γραμματικές χωρίς συμφραζόμενα (context free grammars)
  - Μορφή κανόνων
    - $A \rightarrow a$
    - a: (πιθανός κενή) ακολουθία τερματικών και μη τερματικών συμβόλων
- Επιτρέπουν κανόνες της μορφής:
  - $NP \rightarrow Det\ Nominal$ (δεν επιτρέπεται στις κανονικές γραμματικές)
- Ονομάζονται και Phrase-Structure Grammars (PSG)
  - Ο φορμαλισμός είναι ισοδύναμος με Backus-Naur Form (BNF)

Ιεραρχία γραμματικών του Chomsky (3)

- Τύπος 1: γραμματικές με συμφραζόμενα (context sensitive grammars)
  - Μορφή κανόνων
    - $\alpha A \beta \rightarrow a \gamma \beta$
    - α, β, γ: ακολουθίες τερματικών και μη τερματικών συμβόλων
      - Το γ μη κενό, τα α, β πιθανώς κενά
- Επιτρέπουν κανόνες της μορφής:
  - (Date) → (Day / Month / Year)
    - Αυτή η μορφή δεν επιτρέπεται στις γραμματικές χωρίς συμφραζόμενα

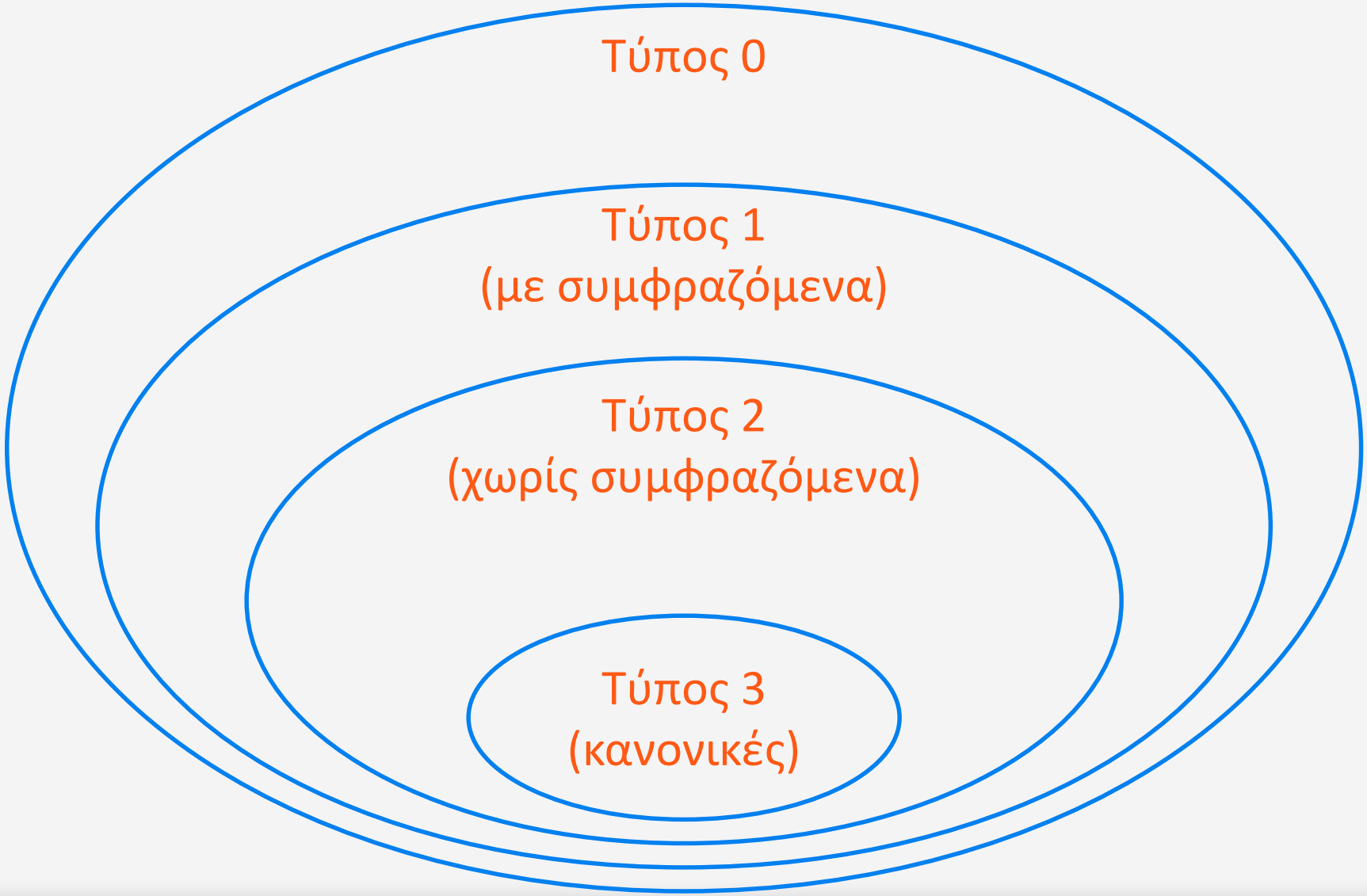
Ιεραρχία γραμματικών του Chomsky (4)

- Τύπος 0: αναδρομικά απαριθμήσιμες
  - Μορφή κανόνων
    - $\alpha \rightarrow \beta$
    - α, β: ακολουθίες τερματικών και μη τερματικών συμβόλων
      - Το α μη κενό

Παραγωγική ισχύς γραμματικών (1)

- Γλώσσες (τύπος 3) $\subset$ γλώσσες (τύπος 2)
  - Π.χ.: οι κανονικές γραμματικές δεν μπορούν να ορίσουν γλώσσες της μορφής $a^n b^n$ (ab, aabb, aaabbb,…)
    - Οι ΓΧΣ μπορούν: $S \to ab, S \to aSb$
- Γλώσσες (τύπος 2) $\subset$ γλώσσες (τύπος 1)
  - Π.χ.: οι κανονικές γραμματικές δεν μπορούν να ορίσουν γλώσσες της μορφής $a^n b^n c^n$
    - Οι ΓΜΣ μπορούν: $S \to abc, S \to aSBc, cB \to Bc, bB \to bb$
- Γλώσσες (τύπος 1) $\subset$ γλώσσες (τύπος 0)

Παραγωγική ισχύς γραμματικών (2)



Μοντέλα υπολογισμού (1)

- Οι κανονικές γραμματικές αντιστοιχούν σε αυτόματα πεπερασμένων καταστάσεων (FSA)
  - Για κάθε κανονική γραμματική, μπορεί να οριστεί FSA που να ορίζει την ίδια ακριβώς γλώσσα (και το αντίστροφο)

Μοντέλα υπολογισμού (2)

- Οι γραμματικές χωρίς συμφραζόμενα αντιστοιχούν σε μη αιτιοκρατικά (non deterministic) FSA με στοίβα
  - Μη αιτιοκρατικό: η τρέχουσα κατάσταση και το σύμβολο εισόδου δεν προσδιορίζουν μονοσήμαντα την επόμενη κατάσταση
- Κάθε μη αιτιοκρατικό FSA μπορεί να μετατραπεί σε αιτιοκρατικό (με περισσότερες καταστάσεις)
  - **Δεν ισχύει** αυτό για αυτόματα με στοίβα

Μοντέλα υπολογισμού (3)

- Οι γραμματικές τύπου 0 αντιστοιχούν σε μηχανές Turing

Τι γραμματικές χρειαζόμαστε; (1)

- Σχεδόν όλα τα συντακτικά φαινόμενα των φυσικών γλωσσών μπορούν να παρασταθούν με **κανονικές** γραμματικές
  - Άρα μπορούμε να κάνουμε συντακτική ανάλυση με αυτόματα πεπερασμένων καταστάσεων
    - Πολύ αποδοτικοί αλγόριθμοι
- Συχνά, όμως, χρησιμοποιούμε ΓΧΣ επειδή είναι πιο σύντομες
  - Και επειδή τα συντακτικά δέντρα που παράγουν είναι πιο χρήσιμα στη σημασιολογική ανάλυση

Τι γραμματικές χρειαζόμαστε; (2)

- Υπάρχουν φαινόμενα για τα οποία φαίνεται να απαιτούνται ΓΧΣ [Jurafsky & Martin 2009]:
  - The cat likes tuna fish.
  - The cat (that) the dog chased likes tuna fish.
- Αντιστοιχία με γλώσσες $a^n b^n$ ($NP^n V^n$ tuna fish)
  - Η τομή (κοινές προτάσεις) των αγγλικών με την κανονική γλώσσα [$NP^n V^m$ tuna fish] είναι η [$NP^n V^n$ tuna fish], που είναι μη κανονική
    - Άρα τα αγγλικά είναι μη κανονική γλώσσα, γιατί η τομή κανονικών γλωσσών είναι κανονική
  - Αλλά και οι άνθρωποι δυσκολεύονται για $n > 2$
  - Για πεπερασμένες τιμές του n αρκούν κανονικές γραμματικές

Τι γραμματικές χρειαζόμαστε; (3)

- Υπάρχουν φαινόμενα σε μερικές γλώσσες που φαίνεται να απαιτούν γραμματικές με συμφραζόμενα
  - Ελβετικά γερμανικά: υπάρχουν εκφράσεις τις μορφής $wa^n b^m c^n d^m y$
- Στις περισσότερες άλλες γλώσσες δεν έχουν βρεθεί τέτοια φαινόμενα

Ανακεφαλαίωση

- Σύνταξη
- Συντακτική ανάλυση
- Συντακτικά δέντρα
- Γραμματικές
- Ιεραρχία γραμματικών Chomsky
- Παραγωγική ισχύ γραμματικών
- Αντιστοιχία με μοντέλα υπολογισμού
- Τύποι γραμματικών για την ΕΦΓ

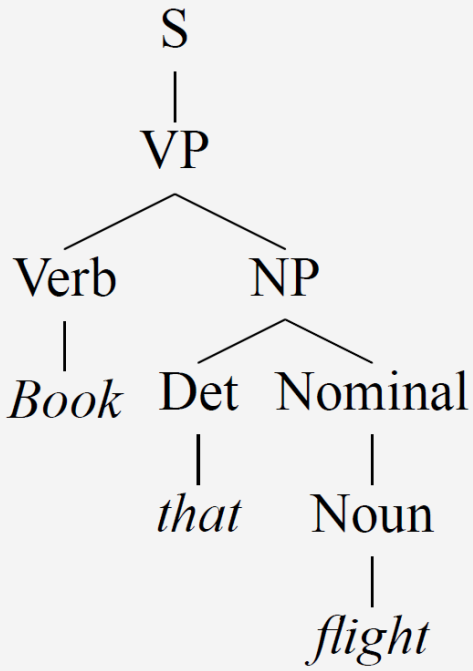
Αλγόριθμοι συντακτικής ανάλυσης

- Είσοδος:
  - Μια γραμματική του τύπου που υποστηρίζει ο αλγόριθμος (π.χ. γραμματική χωρίς συμφραζόμενα)
  - Μια ακολουθία σ από τερματικά σύμβολα της γραμματικής
- Αποκρίσεις:
  - Ανήκει η σ στη γλώσσα που ορίζει η γραμματική;
  - Ποιο είναι το συντακτικό δέντρο της σ;
    - Το συντακτικό δέντρο αποτελεί μια απόδειξη ότι η σ είναι σύμφωνη με τη γραμματική
    - Παρέχει πληροφορίες για τη συντακτική δομή της σ

ΓΧΣ για τμήμα της αγγλικής

$S \rightarrow NP\ VP$
$S \rightarrow Aux\ NP\ VP$
$S \rightarrow VP$
$NP \rightarrow Pronoun$
$NP \rightarrow Proper\text{-}Noun$
$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow Noun$
$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow Verb$
$VP \rightarrow Verb\ NP$
$VP \rightarrow Verb\ NP\ PP$
$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$
$PP \rightarrow Preposition\ NP$

$Det \rightarrow that \mid this \mid a$
$Noun \rightarrow book \mid flight \mid meal \mid money$
$Verb \rightarrow book \mid include \mid prefer$
$Pronoun \rightarrow I \mid she \mid me$
$Proper\text{-}Noun \rightarrow Houston \mid TWA$
$Aux \rightarrow does$
$Preposition \rightarrow from \mid to \mid on \mid near \mid through$



ΣΑ σαν πρόβλημα αναζήτησης

- Αναζήτηση σε αυτόματα πεπερασμένων κατ. (FSA)
  - Εύρεση της σωστής διαδρομής εντός του αυτόματου
  - Ο χώρος αναζήτησης ορίζεται από την δομή του αυτόματου
- Αναζήτηση σε ΓΧΣ
  - Εύρεση του σωστού συντακτικού δέντρου ανάμεσα στα δυνατά συντακτικά δέντρα
  - Ο χώρος αναζήτησης ορίζεται από την γραμματική
- Περιορισμοί (constraints) που προέρχονται:
  - Από την πρόταση εισόδου
  - Αυτόματο/γραμματική

Στρατηγικές αναζήτησης

Δύο στρατηγικές αναζήτησης

- Top-Down
  - Αναζήτηση για δέντρο ξεκινώντας από το “S” (αρχικό σύμβολο), μέχρι να καλυφθούν όλες οι λέξεις της εισόδου
- Bottom-Up
  - Αναζήτηση για δέντρο ξεκινώντας από τις λέξεις, και προσπαθώντας να καλυφθεί το σύμβολο “S”
    - Οι κανόνες εφαρμόζονται αντίστροφα (ταίριασμα δεξιού μέρους)

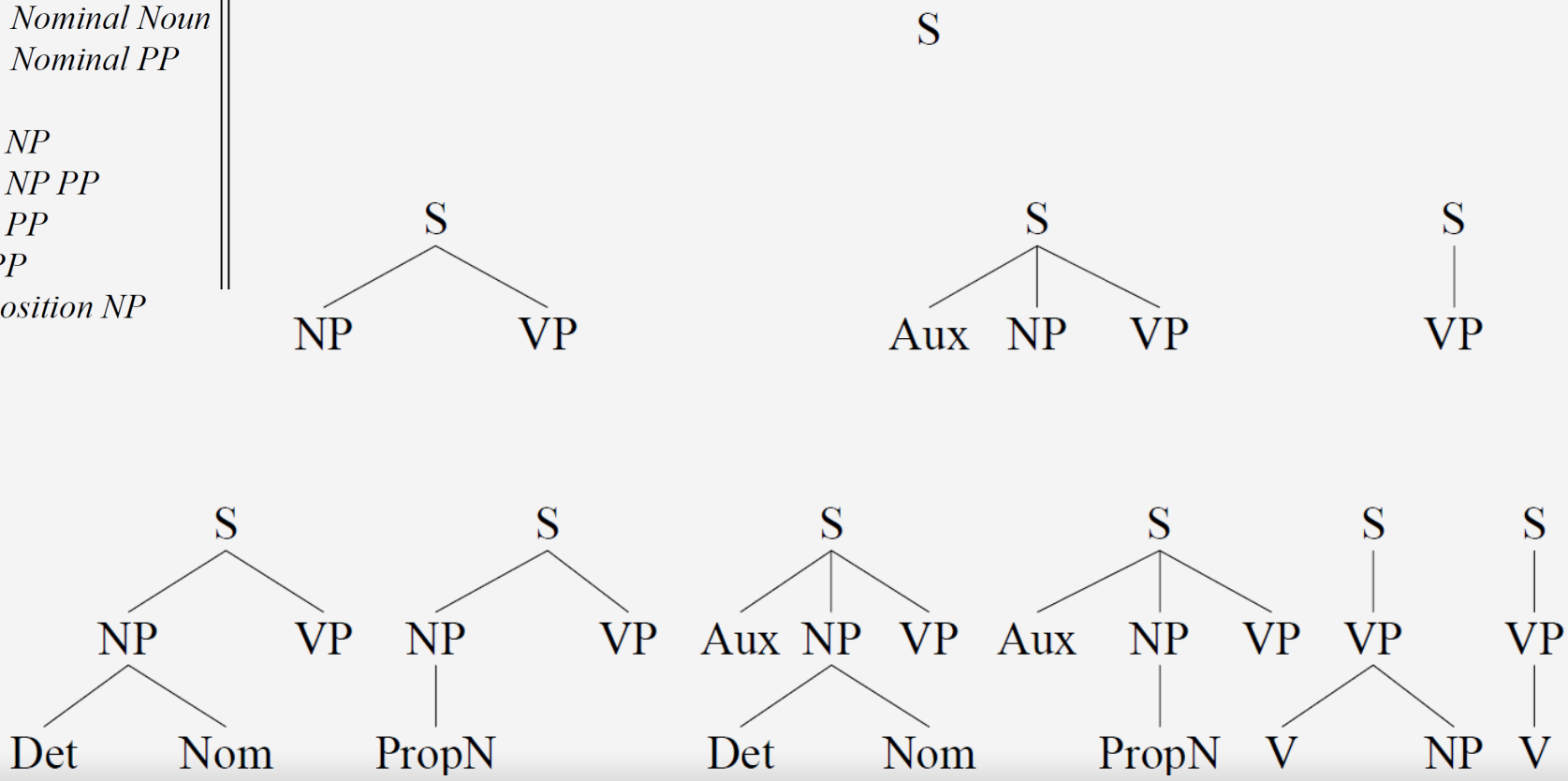
Αναλυτής Top-Down

- Δημιουργεί δέντρα από το αρχικό σύμβολο “S”, προχωρώντας προς τα φύλλα
- Υποθέτοντας την κατασκευή όλως των δέντρων παράλληλα:
  - Εύρεση όλων των δέντρων με ρίζα το “S”
  - Ανάπτυξη όλων των συστατικών (κόμβων) αυτών των δέντρων, μέχρι τα φύλλα
  - Απόρριψη δέντρων που τα φύλλα τους δεν ταιριάζουν με την είσοδο

Χώρος αναζήτησης

$S \rightarrow NP\ VP$
$S \rightarrow Aux\ NP\ VP$
$S \rightarrow VP$
$NP \rightarrow Pronoun$
$NP \rightarrow Proper\text{-}Noun$
$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow Noun$
$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow Verb$
$VP \rightarrow Verb\ NP$
$VP \rightarrow Verb\ NP\ PP$
$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$
$PP \rightarrow Preposition\ NP$

$Det \rightarrow that \mid this \mid a$
$Noun \rightarrow book \mid flight \mid meal \mid money$
$Verb \rightarrow book \mid include \mid prefer$
$Pronoun \rightarrow I \mid she \mid me$
$Proper\text{-}Noun \rightarrow Houston \mid TWA$
$Aux \rightarrow does$
$Preposition \rightarrow from \mid to \mid on \mid near \mid through$

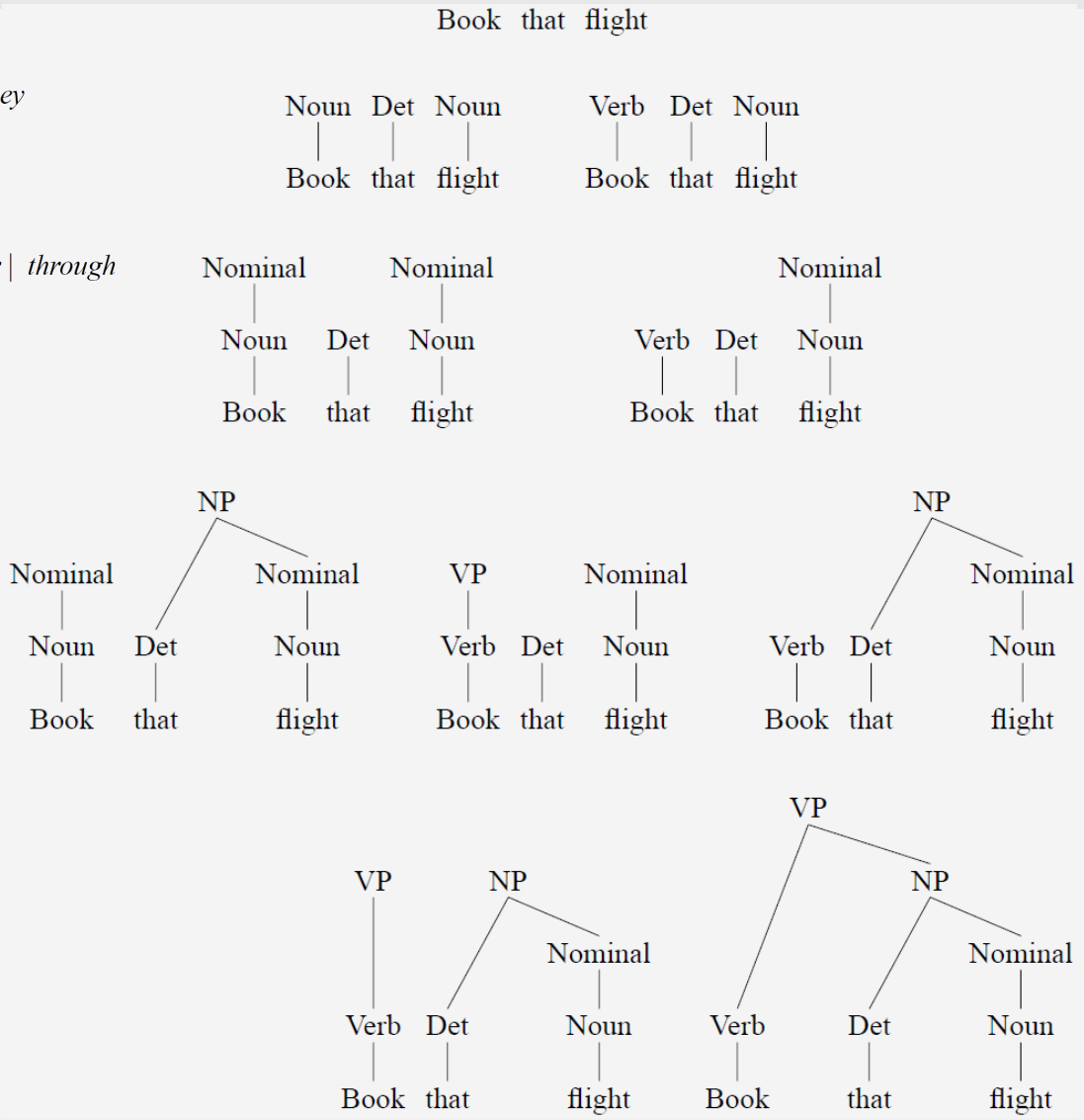


Αναλυτής Bottom-Up

- Ξεκινά από τις λέξεις
- Κατασκευάζει δέντρα εφαρμόζοντας του κανόνες που το δεξί τους μέρος ταιριάζει
- Οδηγούνται από την είσοδο
  - Και όχι από την γραμματική, όπως οι αναλυτές Top-Down

Χώρος αναζήτησης

$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid TWA$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal\ Noun$	
$Nominal \rightarrow Nominal\ PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	
$VP \rightarrow Verb\ NP\ PP$	
$VP \rightarrow Verb\ PP$	
$VP \rightarrow VP\ PP$	
$PP \rightarrow Preposition\ NP$	



Σύγκριση στρατηγικών αναζήτησης (1)

- Αναλυτές Top-Down:
  - Δεν λαμβάνουν υπ’ όψιν άκυρες αναλύσεις (π.χ. δέντρα που δεν περιέχουν το “S”)
  - Χάνουν χρόνο σε δέντρα που δεν ταιριάζουν με την είσοδο
- Αναλυτές Bottom-Up:
  - Δεν λαμβάνουν υπ’ όψιν αναλύσεις που δεν ταιριάζουν με την είσοδο
  - Χάνουν χρόνο παράγοντας άκυρες αναλύσεις

Σύγκριση στρατηγικών αναζήτησης (2)

Κοινό πρόβλημα:

- Πώς πρέπει να γίνει η αναζήτηση στον χώρο των δέντρων;
  - Θα δημιουργηθούν όλα τα εναλλακτικά δέντρα παράλληλα;
  - Ποιος κόμβος πρέπει να αναλυθεί στο επόμενο στάδιο;
  - Ποιος κανόνας πρέπει να εφαρμοστεί στο επόμενο στάδιο;

Στρατηγική και έλεγχος αναζήτησης (1)

- Παραλληλία
  - Εξερεύνηση όλων των δέντρων παράλληλα
- Αναζήτηση σε βάθος (depth first search)
  - Ατζέντα από καταστάσεις: βαθμιαία διαστολή του χώρου αναζήτησης, χρησιμοποιώντας την κατάσταση (δέντρο) που παράχθηκε τελευταία
  - Αν η τρέχουσα κατάσταση είναι ασύμβατη με την είσοδο, οπισθοχώρηση (backtrack) στην πιο πρόσφατη ανεξερεύνητη κατάσταση

Στρατηγική και έλεγχος αναζήτησης (2)

- Ποιος κόμβος πρέπει να αναλυθεί στο επόμενο στάδιο;
  - Αυτός που βρίσκεται «αριστερά»
- Ποιος κανόνας πρέπει να εφαρμοστεί στο επόμενο στάδιο;
  - Ανάλογα με την θέση (σειρά) του στην γραμματική

Βασικός αλγόριθμος

Top-Down, Depth-First, Left-Right

- Αρχικοποίηση ατζέντας με την κατάσταση: δέντρο “S”, δείκτης στην 1η λέξη (cur)
- Επανέλαβε μέχρι: άδεια ατζέντα ή επιτυχής ανάλυση
  - Εφαρμογή όλων των εφαρμόσιμων κανόνων στο αριστερό, μη ανεπτυγμένο κόμβο του cur
    - Αν ο κόμβος είναι τερματικό σύμβολο και ταιριάζει με την είσοδο, «πίεσε» (push) το στην ατζέντα
    - Αλλιώς, «πίεσε» τα νέα δέντρα στην ατζέντα
  - Pop νέο cur από την ατζέντα

Τρία κρίσιμα προβλήματα

- Αριστερή αναδρομή
- Αμφισημία
- Επαναληπτική ανάλυση των ίδιων υπο-δέντρων

Αριστερή αναδρομή

- Η αναζήτηση σε βάθος (depth-first) δεν θα τερματίσει ποτέ, αν η γραμματική περιέχει αριστερή αναδρομή: $A \rightarrow AB\beta$
  - $NP \rightarrow NP\ PP, VP \rightarrow VP\ PP, S \rightarrow S\ \&\ S \rightarrow \in$
  - Έμμεση αναδρομή: $NP \rightarrow Det\ Nominal, Det \rightarrow NP$
- Διαρκείς επαναλήψεις χωρίς κατανάλωση λέξεων εισόδου
  - Με ένα δέντρο που μεγαλώνει διαρκώς

Λύσεις για την αριστερή αναδρομή

- Αλλαγή σειράς κανόνων στην γραμματική
  - $NP \rightarrow NP\ PP, NP \rightarrow Det\ Nominal$
  - $NP \rightarrow Det\ Nominal, NP \rightarrow NP\ PP$
- Απαλοιφή αναδρομικών κανόνων
  - $NP \rightarrow NP\ PP, NP \rightarrow Det\ Nominal$
  - $NP \rightarrow Det\ Nominal\ Stuff,$
    $Stuff \rightarrow PP\ Stuff, Stuff \rightarrow \in$
- Τοποθέτηση (εμπειρικού) ορίου στο βάθος της αναδρομής κατά την ανάλυση
- Αποφυγή αναζήτησης Top-Down

Αμφισημία (1)

- Συντακτικά διφορούμενες προτάσεις
  - «Είδαμε τον επιστήμονα με το τηλεσκόπιο.»
    - Είδαμε [NP τον [Nominal επιστήμονα [PP με το τηλεσκόπιο]]]
    - Όπως «την πτήση από τη Θεσσαλονίκη»
  - «Είδαμε τον επιστήμονα με το τηλεσκόπιο.»
    - Είδαμε [NP τον επιστήμονα] [PP με το τηλεσκόπιο].
    - Θα είχαμε και κανόνα: VP → V NP PP.
  - «Είδαμε τον επιστήμονα με το τηλεσκόπιο από το Παρίσι.»
    - Είδαμε [τον επιστήμονα] [με το τηλεσκόπιο] [από το Παρίσι]
    - Είδαμε [τον επιστήμονα με το τηλεσκόπιο] [από το Παρίσι]
    - Είδαμε [τον επιστήμονα] [με το [τηλεσκόπιο από το Παρίσι]]
    - Είδαμε [τον [επιστήμονα με το [τηλεσκόπιο από το Παρίσι]]]

Αμφισημία (2)

- «Είδαμε τον επιστήμονα με την άσπρη μπλούζα.»
  - Χρειαζόμαστε σημασιολογικούς περιορισμούς που να αποκλείουν την περίπτωση η μπλούζα να είναι το μέσο της παρατήρησης
- Από καθαρά συντακτική σκοπιά, οι περισσότερες προτάσεις είναι εξαιρετικά διφορούμενες
  - Πολύ μεγάλος αριθμός συντακτικών δένδρων (συχνά εκθετική αύξηση όσο αυξάνει ο αριθμός των φράσεων που συνδυάζονται)
  - Χρονοβόρο να ανακαλύψουμε και να επιστρέψουμε όλα τα συντακτικά δέντρα ξεχωριστά
  - Πρόβλημα για όλους τους απλούς αλγορίθμους συντακτικής ανάλυσης που έχουμε εξετάσει ως τώρα

Κανονική μορφή Chomsky

Γραμματικές χωρίς συμφραζόμενα σε κανονική μορφή Chomsky (CNF)

- Επιτρέπονται μόνο κανόνες της μορφής $A \rightarrow BC$ και $A \rightarrow w$, όπου A, B, C μη τερματικά και w τερματικό

- Κάθε ΓΧΣ μπορεί να μετατραπεί σε CNF
  - Χωρίς να σημαίνει ότι τα συντακτικά δέντρα παραμένουν ίδια
- Γραμματικές σε CNF μπορούν να αναλυθούν με τον αλγόριθμο CKY (Cocke-Younger-Kasami, 1960)
  - Αλγόριθμος δυναμικού προγραμματισμού

Δυναμικός προγραμματισμός

- Δημιουργία πινάκων με λύσεις σε υπο-προβλήματα (π.χ. υπο-δέντρα), καθώς γίνεται η ανάλυση
- Αναζήτηση έτοιμων λύσεων αντί για την επανα-ανάλυσή τους
- Όλα τα δέντρα αποθηκεύονται έμμεσα
  - Είναι διαθέσιμα για αποσαφήνιση σε μετέπειτα στάδιο

Ο αλγόριθμος CKY (1)

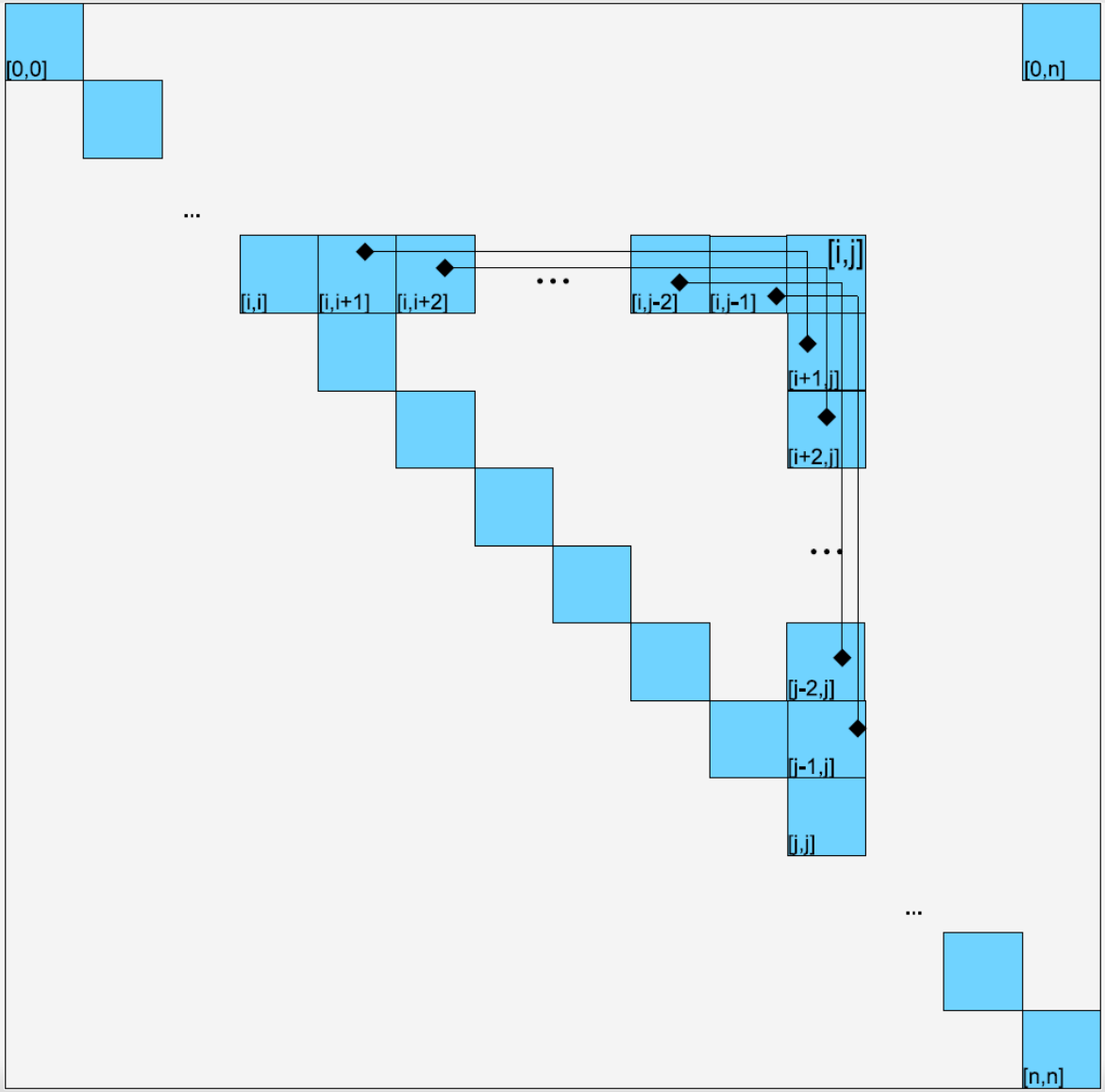
- Γραμματική σε CNF
  - Κάθε κόμβος, έχει το πολύ 2 παιδιά
  - Ένας δισδιάστατος πίνακας μπορεί να αναπαραστήσει ένα δέντρο
    - Για είσοδο ν λέξεων, χρειαζόμαστε $(\nu + 1) * (\nu + 1)$
  - Κάθε κελί $[i, j]$ περιέχει το σύνολο των μη τερματικών συμβόλων που περιέχουν την είσοδο από την λέξη i μέχρι την λέξη j
    - Ξεκινώντας από το 0, σημαδεύουμε τα κενά μεταξύ των λέξεων

Ο αλγόριθμος CKY (2)

function CKY-PARSE(*words*, *grammar*) **returns** *table*

for $j \leftarrow$ **from** 1 **to** LENGTH(*words*) **do**
  $table[j-1, j] \leftarrow \{A \mid A \rightarrow words[j] \in grammar\}$
  for $i \leftarrow$ **from** $j-2$ **downto** 0 **do**
    for $k \leftarrow i+1$ **to** $j-1$ **do**
      $table[i,j] \leftarrow table[i,j] \cup$
        $\{A \mid A \rightarrow BC \in grammar,$
          $B \in table[i,k],$
          $C \in table[k,j]\}$

Ο αλγόριθμος CKY (3)



Μια απλή γραμματική

- S → V NP
- V → θέλω, V → επιθυμώ
- NP → Det Nominal
- Nominal → Adj Nominal
- Det → μια
- Adj → πρωινή, Adj → απογευματινή
- N → πτήση
- Nominal → πτήση

Ο αλγόριθμος CKY

0 θέλω 1 μια 2 πρωινή 3 πτήση 4

	0	1	2	3	4
0		V (0,1)			
1			Det (1,2)		
2				Adj (2,3)	
3					Nominal N (3,4)

Ο αλγόριθμος CKY

0 θέλω 1 μια 2 πρωινή 3 πτήση 4

	0	1	2	3	4
0		V (0,1)	X (0,2)		
1			Det (1,2)		
2				Adj (2,3)	
3					Nominal N (3,4)

Δεν υπάρχει κανόνας που να συνδυάζει V με Det

Ο αλγόριθμος CKY

0 θέλω 1 μια 2 πρωινή 3 πτήση 4

	0	1	2	3	4
0		V (0,1)	(0,2)		
1			Det (1,2)	X (1,3)	
2				Adj (2,3)	
3					Nominal N (3,4)

Δεν υπάρχει κανόνας που να συνδυάζει Det με Adj

Ο αλγόριθμος CKY

0 θέλω 1 μια 2 πρωινή 3 πτήση 4

	0	1	2	3	4
0		V (0,1)	(0,2)	X (0,3)	
1			Det (1,2)	(1,3)	
2				Adj (2,3)	
3					Nominal N (3,4)

Το (1,3) είναι κενό

Ο αλγόριθμος CKY

0 θέλω 1 μια 2 πρωινή 3 πτήση 4

	0	1	2	3	4
0		V (0,1)	(0,2)	X (0,3)	
1			Det (1,2)	(1,3)	
2				Adj (2,3)	
3					Nominal N (3,4)

Το (0,2) είναι κενό

Ο αλγόριθμος CKY

0 θέλω 1 μια 2 πρωινή 3 πτήση 4

	0	1	2	3	4
0		V (0,1)	(0,2)	(0,3)	
1			Det (1,2)	(1,3)	
2				Adj (2,3)	Nominal (2,4)
3					Nominal N (3,4)

Ο αλγόριθμος CKY

0 θέλω 1 μια 2 πρωινή 3 πτήση 4

	0	1	2	3	4
0		V (0,1)	 (0,2)	 (0,3)	
1			Det (1,2)	 (1,3)	NP, X (1,4)
2				Adj (2,3)	Nominal (2,4)
3					Nominal N (3,4)

Ο αλγόριθμος CKY

0 θέλω 1 μια 2 πρωινή 3 πτήση 4

	0	1	2	3	4
0		V (0,1)	(0,2)	(0,3)	S, X, X (0,4)
1			Det (1,2)	(1,3)	NP (1,4)
2				Adj (2,3)	Nominal (2,4)
3					Nominal N (3,4)