Εργαστήριο επισημείωσης δένδρων εξάρτησης

Προκόπης Προκοπίδης

Ινστιτούτο Επεξεργασίας του Λόγου/Ε.Κ. Αθηνά

10 Ιουνίου 2014

ΜΠΣ Τεχνογλωσσία: Επεξεργασία Σωμάτων Κειμένων

Εισαγωγή

- Γενικότερος στόχος: ανάπτυξη και αξιολόγηση εφαρμογών Επεξεργασίας Φυσικής Γλώσσας
  - εξαγωγή πληροφορίας
  - μηχανική μετάφραση
  - αυτόματη περίληψη, συντόμευση κειμένου κ.α.
- Προϋπόθεση: Ανάπτυξη αυτόματων συντακτικών αναλυτών (parsers) που
  - δέχονται αυτόματα επισημειωμένη είσοδο (στο επίπεδο της λέξης)
  - αναπαριστούν κάθε πρόταση της εισόδου ως ένα συντακτικό δέντρο
  - εκπαιδεύονται και αξιολογούνται σε χειρωνακτικά επισημειωμένες δενδροτράπεζες (treebanks)
  - συχνά χρησιμοποιούν αναπαραστάσεις βασισμένες σε δένδρα εξαρτήσεων (dependency trees)

Περιεχόμενα

Λεκτικός αναλυτής και αναγνωριστής προτάσεων

- Λεκτική ανάλυση κειμένου με γραμματική κανονικών εκφράσεων και λίστες συντμήσεων
- Έξοδος
  - όρια παραγράφων και προτάσεων
  - όρια λεκτικών μονάδων
  - επισήμανση αρχικών, συντμήσεων, αρκτικόλεξων ...

Παράδειγμα

... του Κ.Κ.Ε. Η παρούσα απόφαση εστιάζει στην παρουσίαση ...
... συγκέντρωση της Νεολαίας ΠΑ.ΣΟ.Κ. Νοτίων Προαστίων ...

Μορφολογικός χαρακτηριστής

- Χαρακτηρίζει κάθε λέξη ως προς το μέρος του λόγου στο οποίο ανήκει
- Ανάλογα με το μέρος του λόγου, αποδίδει επίσης πληροφορίες υποκατηγοριοποίησης, όπως, π.χ.,
  - αριθμό, γένος και πτώση για τα ουσιαστικά
  - φωνή, χρόνο και πρόσωπο για τα ρήματα

Παράδειγμα

λόγος → NoCmMaSgNm
No = ουσιαστικό, Cm = κοινό, Ma = αρσενικό, Sg = ενικός αριθμός, Nm = ονομαστική

Λημματοποιητής

- Δέχεται ως είσοδο μια λέξη μαζί με την πληροφορία από τον μορφολογικό χαρακτηριστή
- Με την υποστήριξη λεξικών επιστρέφει το λήμμα

Παράδειγμα

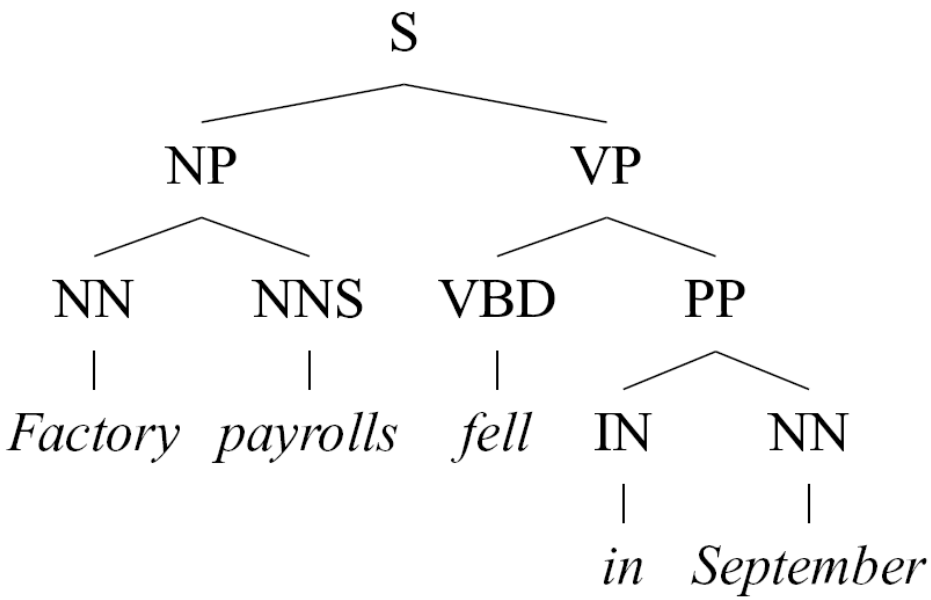
(οι, τις) κρατήσεις/NoCmFePlNm, Ac] → κράτηση
(να) κρατήσεις/VbMnIdXx02SgXxPeAvXx → κρατάω

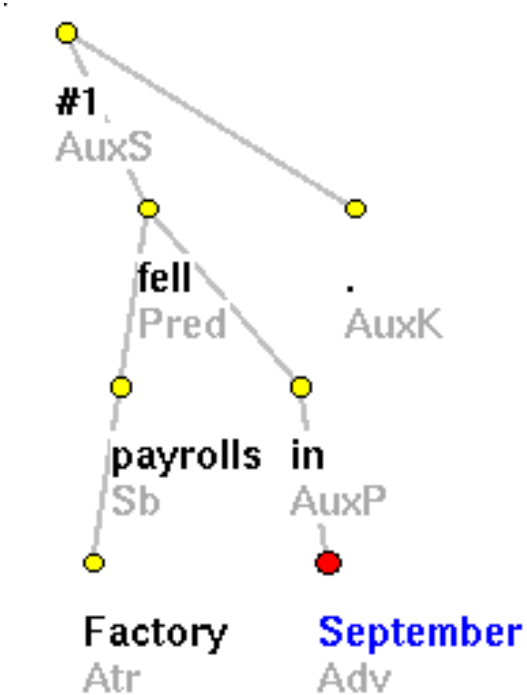
- Περισσότερες πληροφορίες: [PGP11]

Αναπαραστάσεις με δένδρα εξαρτήσεων

- Περισσότερο μια οικογένεια από αναπαραστάσεις, παρά μια αυστηρά κοινή θεώρηση
- Βασική παραδοχή: η συντακτική δομή αποτελείται από λεκτικές μονάδες
- Οι μονάδες αυτές συνδέονται με δυαδικές ασυμμετρικές σχέσεις που ονομάζονται εξαρτήσεις (dependencies) από έναν εξαρτώμενο κόμβο σε έναν κόμβο-κεφαλή
- Σχέσεις γνωστές από την παραδοσιακή γραμματική (υποκείμενο, αντικείμενο, κ.α.)
- Απουσία φραστικών κόμβων όπως στα Δένδρα Φραστικής Δομής

Παράδειγμα αναπαραστάσεων με ΔΦΔ και ΔΕ





Γιατί η ανάλυση με ΔΕ; Σε τι χρησιμεύει;

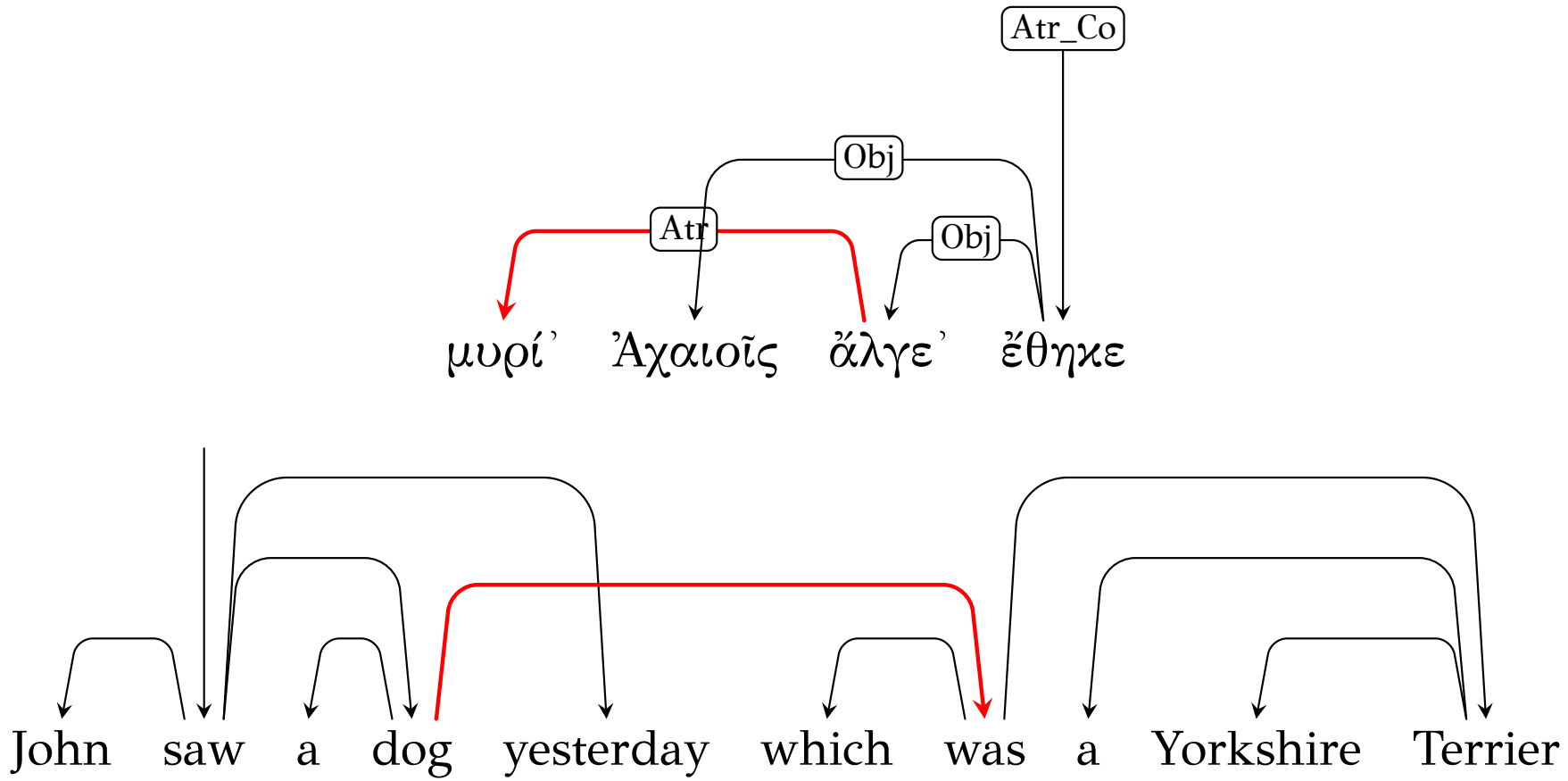
- Τα δένδρα εξαρτήσεων
  - Είναι υπολογιστικά αποδοτικός τρόπος αναπαράστασης:
    - Αριθμός κόμβων ίσος με αριθμό λέξεων
    - Απουσία κενών κόμβων
  - Προσφέρουν μια κατάλληλη μοντελοποίηση για γλώσσες με ελεύθερη σειρά όρων
  - Ένα βήμα πιο μπροστά σε σχέση με τη σημασιολογία
- Μεγάλη ανάπτυξη τα τελευταία χρόνια στην έρευνα για ΣΑ με ΔΕ

Διαφορές στις αναπαραστάσεις

- Διαφορετικά σύνολα με ετικέτες για τις σχέσεις
- Διαφορετικές κατηγορίες ετικετών
  - Συντακτικές κατηγορίες
  - Θεματικοί ρόλοι
- Άλλες διαφορές. Π.χ. στη σύνδεση κατά παράταξη, κεφαλή μπορεί να είναι
  - Ο παρατακτικός σύνδεσμος
  - Ο πρώτος όρος (με τον σύνδεσμο ως εξαρτώμενο κόμβο, και τον δεύτερο όρο ως εξαρτώμενο κόμβο από το σύνδεσμο)

Προβολικότητα

- Προβολικό δένδρο: για κάθε ακμή από μία λέξη-κεφαλή *w* προς μία άλλη λέξη *u*, ισχύει ότι η *w* είναι πρόγονος κάθε λέξης που μεσολαβεί μέσα στην πρόταση ανάμεσα στην *w* και την *u*



- Μη-προβολικά δένδρα σε πολλές δενδροτράπεζες με ΔΕ

Prague Dependency Treebank 3.0

- DT για τα Τσεχικά: http://ufal.mff.cuni.cz/pdt3.0/
- Πολυεπίπεδος σχολιασμός
  - Μορφολογικό επίπεδο
  - Αναλυτικό επίπεδο (28 βασικές ΣΕ)
  - Σημασιολογικοί ρόλοι, συναναφορά, κ.α.
- 2 M tokens από άρθρα σε εφημερίδες και περιοδικά, διορθωμένα στο επίπεδο της μορφολογίας
- $>$ 1.5 M tokens (87K προτάσεις) στο αναλυτικό και ...
- $>$ 0.8 M tokens στο σημασιολογικό επίπεδο
- Ανάπτυξη εργαλείων χειρωνακτικής επισημείωσης
- Προβολικότητα
  - Μόνο το 1.81% των ακμών του PDT είναι μη-προβολικές αλλά ...
  - το 23.15% των προτάσεων περιέχει τουλάχιστον μία

Στοιχεία για ορισμένα άλλα DTs

Γλώσσα	**Προτάσεις**	**Λέξεις**	**Μέσο μήκ. πρότ.**	**Nproj**
Arabic	3.043	116.793	38.38	0.37
Basque	11.226	151.604	13.50	1.27
Danish	5.512	100.238	18.19	0.99
GDT 2007	2.902	70.223	24.20	1.17
Greek Anc.	21.160	308.882	14.60	19.58
Russian	34.985	497.465	14.26	0.83
Turkish	5.935	69.695	11.74	5.33
English	40.613	991.535	24.41	0.39
German	38.020	680.710	17.90	2.33
Spanish	15.984	477.810	29.89	0.00

Πηγή: [Zem+12]

- Πρόσφατες προσπάθειες για την ομογενοποίηση αυτών των πόρων ([Zem+12], [McD+13])

Περιεχόμενα

Βασικά στοιχεία για το Greek Dependency Treebank

- Χειρωνακτική επισημείωση στα συντακτικό επίπεδο με δέντρα εξαρτήσεων
- Χειρωνακτική επισημείωση στα επίπεδα της μορφολογίας και του λήμματος
- Επισημειωτές: κυρίως ερευνητές του ΙΕΛ, μεταπτυχιακές/οί του ΜΠΣ Τεχνογλωσσία, τελειόφοιτες/οι του Τμήματος Γλωσσολογίας του ΕΚΠΑ
- Συνεχής προσπάθεια για αύξηση του υλικού μέχρι σήμερα, στα πλαίσια ερευνητικών προγραμμάτων
- 21827 μοναδικοί τύποι, 11005 λήμματα

Προτάσεις	**Λέξεις**	**Tokens**	**Μέσ. μήκος**	**Κείμενα**
5676	116662	130903	23.07	249

Προέλευση των κειμένων του GDT

Χρονολογία	**% προτάσ.**
1993-99	22.97
2000-09	71.19
2010-σήμερα	5.83
Πηγή	**% προτάσ.**
Ειδησεογραφικά sites	50.11
Ομιλίες από το ευρωκοινοβούλιο	22.92
Τουριστικές εκπομπές	16.95
Ελληνικά wikipedia & wikinews	10.02

Σχήμα επισημείωσης

- Οδηγίες βασισμένες στο σχήμα του PDT, προσαρμοσμένες στη NE
- 18 βασικές σχέσεις με παραλλαγές για τις περιπτώσεις της σύνδεσης κατά παράταξης, της παράθεσης κ.α.

Είδη εξαρτήσεων στο GDT (I)

Afun	Περιγραφή
AuxS	Η ρίζα κάθε προτασιακού δέντρου
Pred	Το κατηγόρημα της κύριας πρότασης μιας περιόδου
Sb	Υποκείμενο
Obj	Άμεσο Αντικείμενο
IObj	Έμμεσο Αντικείμενο
Pnom	Κόμβος που εξαρτάται από το ρήμα και έχει ρόλο κατηγορούμενου του υποκειμένου ή του αντικειμένου του ρήματος
Atr	Προσδιορισμοί του Ονόματος
AuxP	Πρόθεση
AuxC	Υποτακτικός σύνδεσμος

Είδη εξαρτήσεων στο GDT (II)

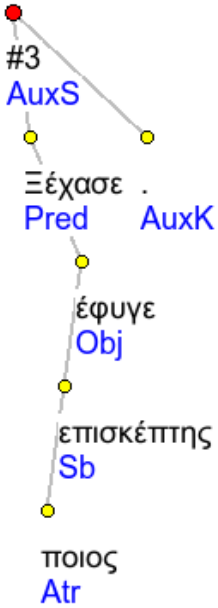
Afun	Περιγραφή
Coord	Κόμβος που κυβερνά κόμβους σε παρατακτική σύνδεση
Apos	Κόμβος που κυβερνά κόμβους σε παράθεση
*_Co	Η σχέση ενός στοιχείου της παρατακτικής σύνδεσης (π.χ. Sb_Co, Obj_Co κ.λπ.) με τον κόμβο που κυβερνά τον κόμβο Coord
*_Ap	Η σχέση ενός στοιχείου της παράθεσης (π.χ. Sb_Ap, Obj_Ap κ.λπ.) με τον κόμβο που κυβερνά τον κόμβο Apos
*_Pa	Η σχέση της κεφαλής μιας παρενθετικής δομής (π.χ. Adv_Pa, Atr_Pa κ.λπ.) με τον κόμβο που την κυβερνά
AuxX	Κόμμα
AuxK	Τερματικά σημεία στίξης
AuxG	Άλλα, μη τερματικά, σημεία στίξης
ExD	Αποδίδεται στις λέξεις των οποίων ο κυβερνών κόμβος ελλείπει (Externally-Dependent)
AuxY	Άλλα βοηθητικά στοιχεία της πρότασης

Διαδικασία επισημείωσης σχέσεων εξάρτησης

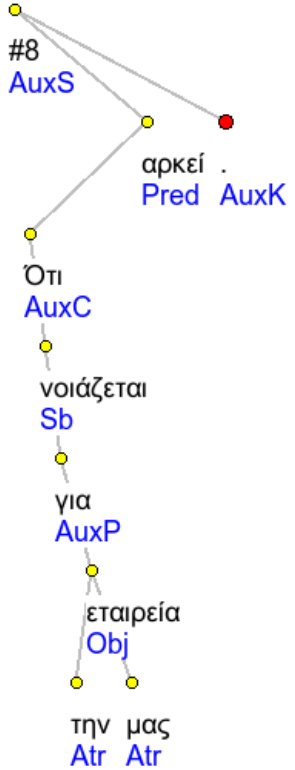
- Η ρίζα του δέντρου είναι ένας τεχνητός κόμβος με την τιμή AuxS.
- Σε μια πρόταση με πρωτοτυπική δομή, το κατηγόρημα (Pred) της κύριας πρότασης εξαρτάται από τη ρίζα του δέντρου (AuxS).
- Σε μια πρόταση όπου τα κατηγορήματα συνδέονται παρατακτικά, από τη ρίζα του δέντρου εξαρτάται ο κύριος κόμβος της παρατακτικής σύνδεσης (Coord).
- Για να αναλύσουμε τις εξαρτήσεις της πρότασης:
  - εντοπίζουμε το υποκείμενο και το προσαρτούμε από το ρήμα της κύριας, αποδίδοντάς του το χαρακτηρισμό Sb.
  - συνεχίζουμε με τα υπόλοιπα συμπληρώματα του ρήματος (Obj, IObj, κ.λπ),
  - χαρακτηρίζουμε τα στοιχεία που προσδιορίζουν ή τροποποιούν τα ονόματα και τα ρήματα (Atr, Adv, κ.λπ.) Προσαρτούμε όλες τις υπόλοιπες λέξεις και σημεία στίξης.

Κατηγόρημα - Υποκείμενο

Ξέχασε πότε έφυγε.

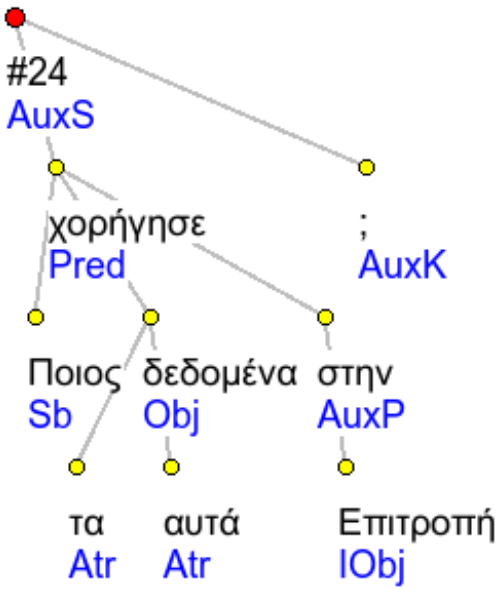


Ξέχασε ποιος επισκέπτης έφυγε.

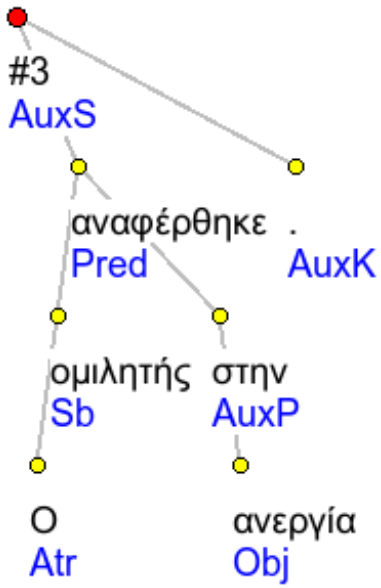


Ότι νοιάζεται για την εταιρεία μας αρκεί.

Αντικείμενο - Βοηθητικοί κόμβοι



Ποιος χορήγησε τα δεδομένα αυτά στην Επιτροπή;



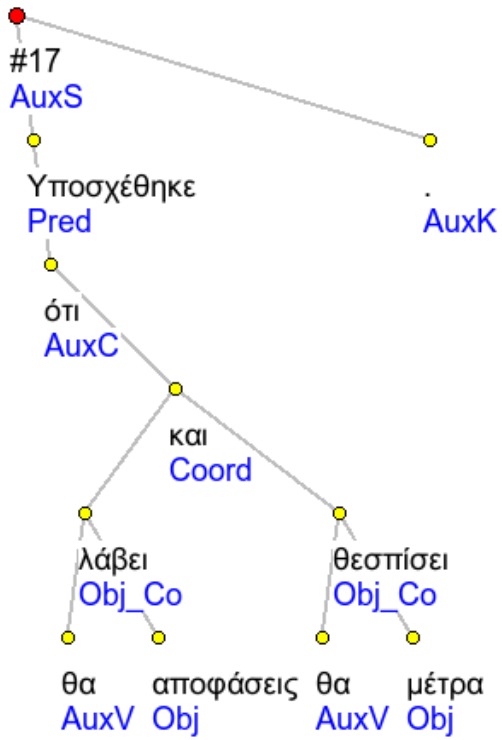
Ο ομιλητής αναφέρθηκε στην ανεργία.

#19 Ο εισηγητής εκτιμά πως υπάρχει περιθώριο.

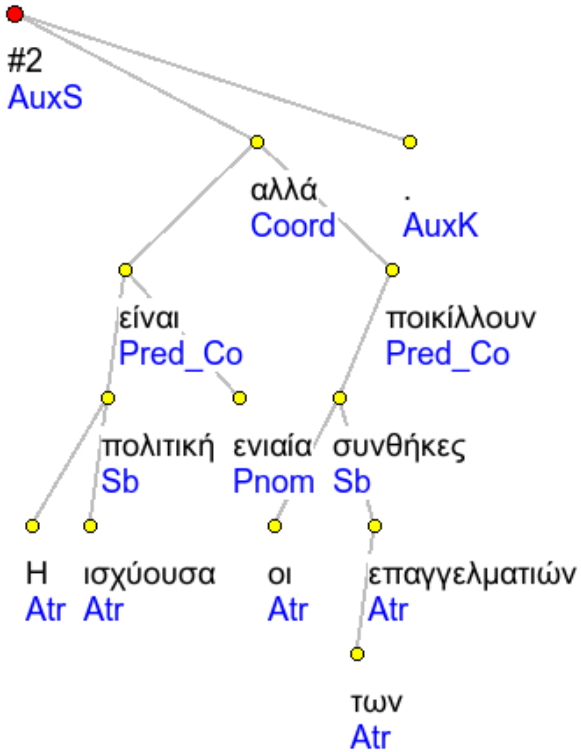
Διαδικασία επισημείωσης σχέσεων εξάρτησης

- Τα προβλήματα της παρατακτικής σύνδεσης και της παράθεσης επιλύονται τεχνικά με την επιλογή ενός «κύριου» κόμβου (ενός παρατακτικού συνδέσμου ή ενός κόμματος). Τα μέλη των δομών αυτών εξαρτώνται από αυτό τον κόμβο και χαρακτηρίζονται με ένα ιδιαίτερο επίθημα (*_Co ή *_Ap).
- Σε περιπτώσεις έλλειψης και όταν απουσιάζει η κεφαλή ενός κόμβου, αποδίδουμε στον τελευταίο την ειδική τιμή ExD.

Παράταξη

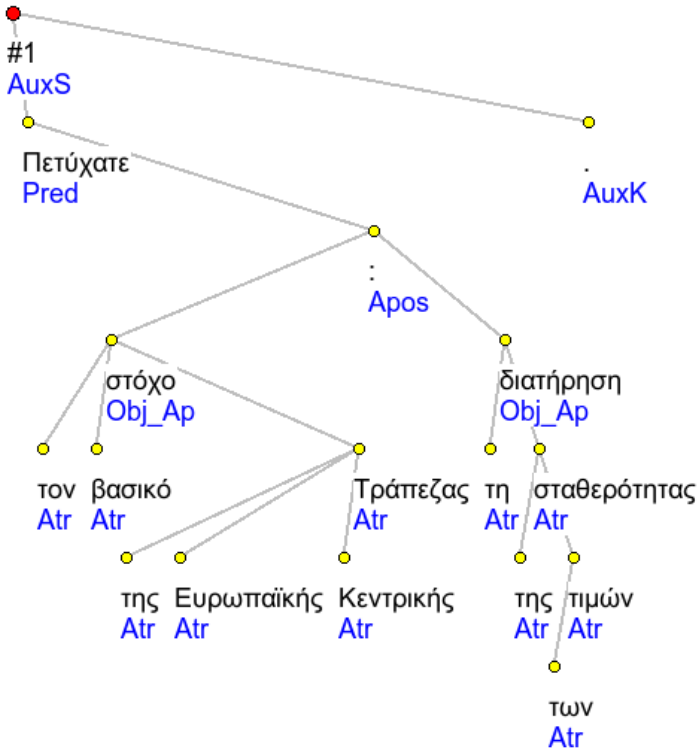


Υποσχέθηκε ότι θα λάβει υποσχέσεις και θα θεσπίσει μέτρα.

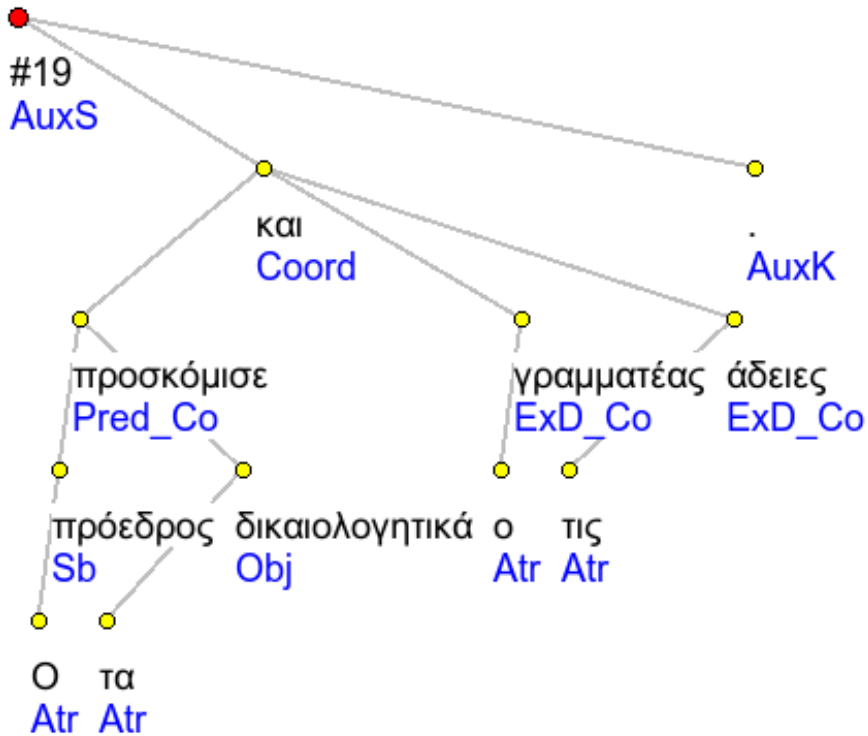


Η ισχύουσα πολιτική είναι ενιαία αλλά οι συνθήκες των επαγγελματιών ποικίλλουν.

Παράθεση - Έλλειψη



Πετύχατε τον βασικό στόχο της Ευρωπαϊκής Κεντρικής Τράπεζας: τη διατήρηση της σταθερότητας των τιμών.



Ο πρόεδρος προσκόμισε τα δικαιολογητικά και ο γραμματέας τις άδειες.

Περιεχόμενα

Συντακτικοί αναλυτές με ΔΕ

- Είσοδος: λέξεις, μορφολογική πληροφορία και λήμματα
- Έξοδος: δένδρα εξαρτήσεων στα οποία έχει αναγνωριστεί για κάθε λέξη η λέξη-κεφαλή της και η σχέση με την κεφαλή
- Μετρικές αξιολόγησης
  - Labeled Attachment Score (LAS) = το ποσοστό των λέξεων για τις οποίες αναγνωρίζεται η σωστή κεφαλή και η σωστή σχέση εξάρτησης
  - Unlabeled Attachment Score (UAS) = το ποσοστό των λέξεων για τις οποίες αναγνωρίζεται η σωστή κεφαλή
  - Label Accuracy (LACC) = το ποσοστό των λέξεων για τις οποίες αναγνωρίζεται η σωστή σχέση εξάρτησης

ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

$$[\text{ROOT}]_S\ [\text{Ηρθε, πολύς, κόσμος, χτες}]_Q$$

ROOT Ήρθε πολύς κόσμος χτες

ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

$$[\text{ROOT}, \text{Ήρθε}]_S \; [\text{πολύς}, \text{κόσμος}, \text{χτες}]_Q \; \textbf{SHIFT}$$

$$\text{ROOT} \quad \text{Ήρθε} \quad \text{πολύς} \quad \text{κόσμος} \quad \text{χτες}$$

ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

$$[\text{ROOT}, \text{Ήρθε}, \text{πολύς}]_S \; [\text{κόσμος}, \text{χτες}]_Q \; \textbf{SHIFT}$$

$$\text{ROOT} \quad \text{Ήρθε} \quad \text{πολύς} \quad \text{κόσμος} \quad \text{χτες}$$

ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

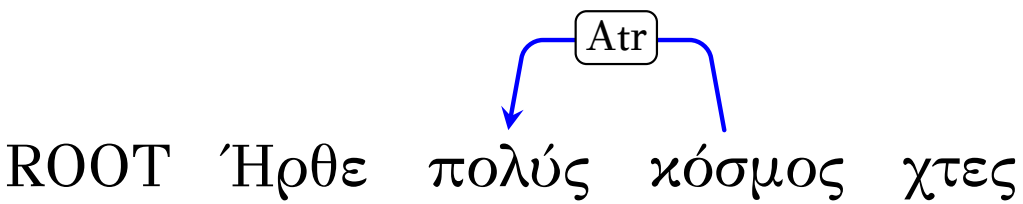
$$[\text{ROOT}, \text{Ήρθε}, \text{πολύς}, \text{κόσμος}]_S \; [\text{χτες}]_Q \; \textbf{SHIFT}$$

$$\text{ROOT} \quad \text{Ήρθε} \quad \text{πολύς} \quad \text{κόσμος} \quad \text{χτες}$$

ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

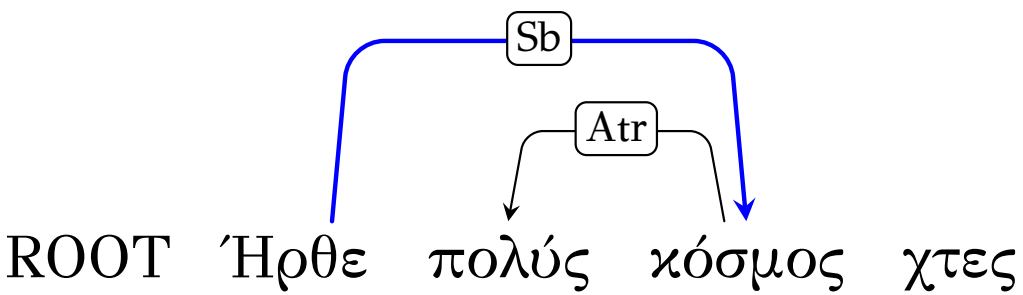
$$[\text{ROOT}, \text{Ήρθε}, \text{κόσμος}]_S\ [\text{χτες}]_Q\ \textbf{LEFT\_ARC}$$



ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

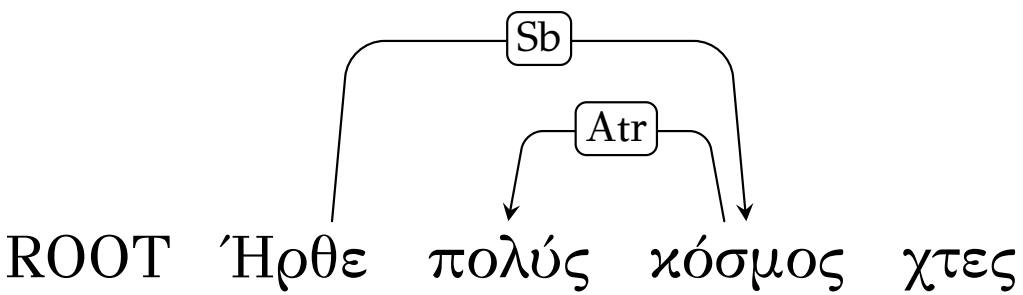
$$[\text{ROOT}, \text{Ήρθε}]_S \; [\text{χτες}]_Q \; \textbf{RIGHT\_ARC}$$



ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

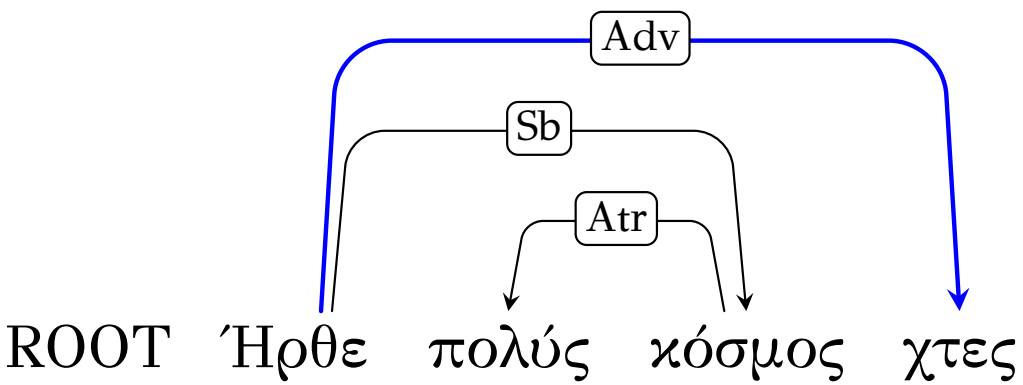
$$[\text{ROOT}, \text{Ήρθε}, \text{χτες}]_S\ []_Q\ \textbf{SHIFT}$$



ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

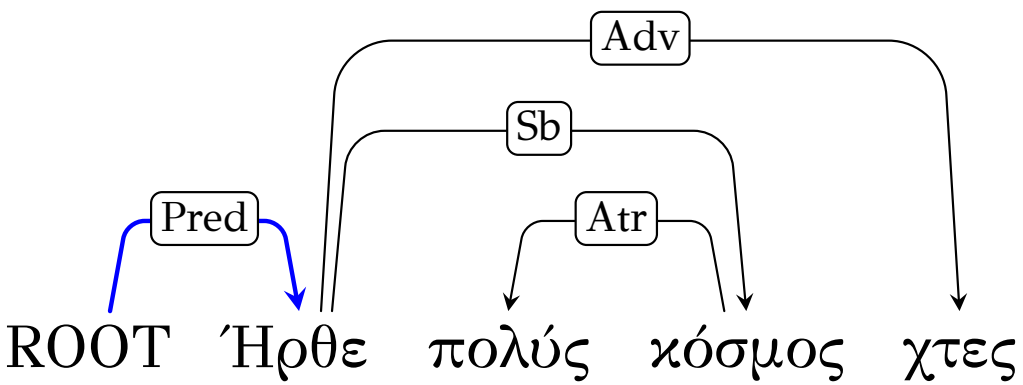
$$[\text{ROOT}, \text{Ήρθε}]_S\ []_Q\ \textbf{RIGHT\_ARC}$$



ΣΑ βασισμένοι στις μεταβάσεις

- Transition-based parsers
- Διατρέχουν την είσοδο από αριστερά προς τα δεξιά
- Δομές δεδομένων: στοίβα με εξετασθείσες λέξεις, ουρά με την υπολειπόμενη είσοδο
- Στόχος η εύρεση μιας ακολουθίας μεταβάσεων από την αρχική κατάσταση της ανάλυσης σε μια τελική κατάσταση:

$$[\text{ROOT}]_S\ []_Q\ \textbf{RIGHT\_ARC}$$



Χαρακτηριστικά

- Οι αποφάσεις αυτών των αναλυτών επηρεάζονται από τοπικά χαρακτηριστικά όπως, π.χ. ...
- ... οι ίδιες οι λέξεις στην κορυφή της στοίβας και στην αρχή της λίστας (S.LEX, Q.LEX)
- το μέρος του λόγου (S.POS, Q.POS)
- το μέρος του λόγου των επόμενων τριών λέξεων στην είσοδο, και
- εφόσον είναι διαθέσιμες,
  - η σχέση του S προς τον αριστερότερο και δεξιότερο εξαρτώμενο κόμβο του (SL.DEP, SR.DEP), μαζί
  - με τα αντίστοιχα μέρη του λόγου (SL.POS, SR.POS) κ.α.

Χαρακτηριστικά αποτελέσματα αξιολόγησης για διάφορες γλώσσες

- CoNLL shared task 2007 [Niv+07]:
  - $84 \leq \text{LAS} \leq 90$: Catalan, Chinese, English, Italian
  - $76 \leq \text{LAS} \leq 80$: Arabic, Basque, Czech, Greek, Hungarian, Turkish
- Οι γλώσσες με ελεύθερη σειρά όρων και πλουσιότερη μορφολογία μοιάζει να παρουσιάζουν χαμηλότερα αποτελέσματα
- Σήμερα, το LAS για τα Αγγλικά $\approx 90\%$

Περιεχόμενα

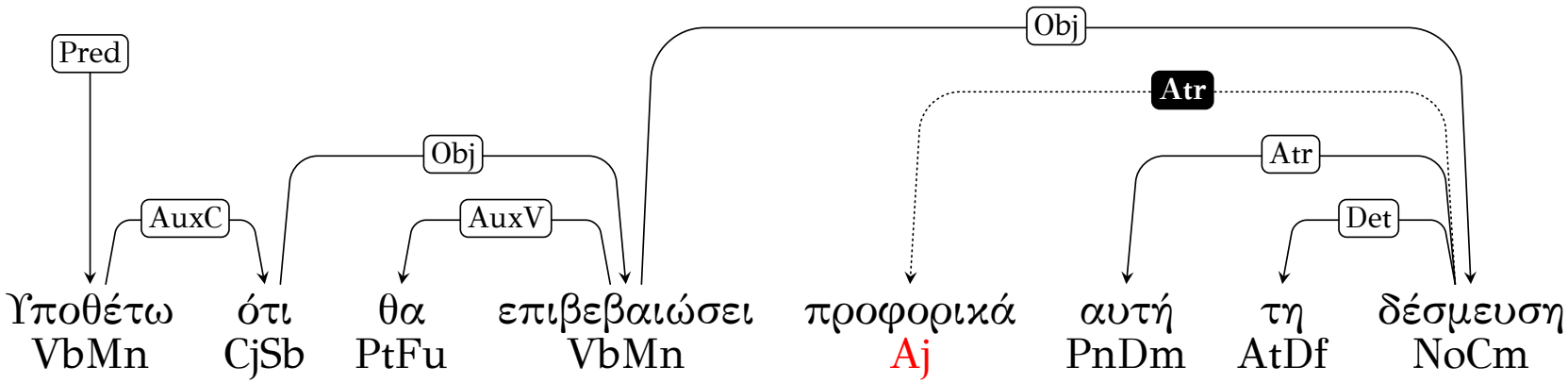
Αποτελέσματα από την εκπαίδευση ΣΑ για την ΝΕ

- Σώμα εκπαίδευσης/αξιολόγησης: 90/10% των προτάσεων του GDT

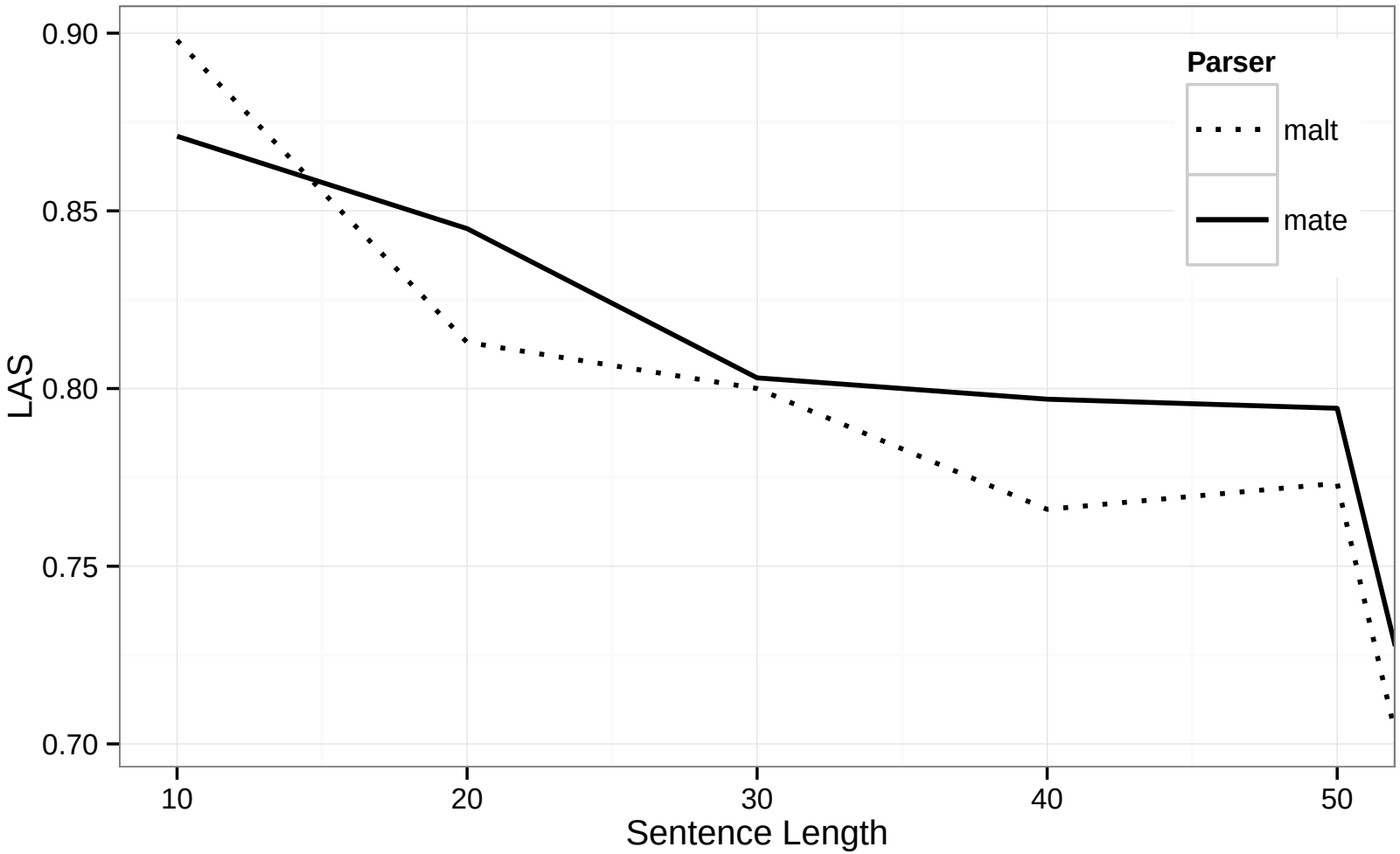
	MPL	**APL**	**APML**	**MPAL**
LAS	79.74	76.29	76.40	79.68
UAS	85.83	83.57	83.69	85.77
LACC	87.94	85.67	85.72	87.91

Πίνακας: Results from parsing GDT: MPL refers to training and testing on manually validated POS, morphological features and lemmas; APL is evaluation on automatic POS, features and lemmas; APML is evaluation on automatic morphology and gold lemmas; MPAL on gold morphology and automatic lemmas.

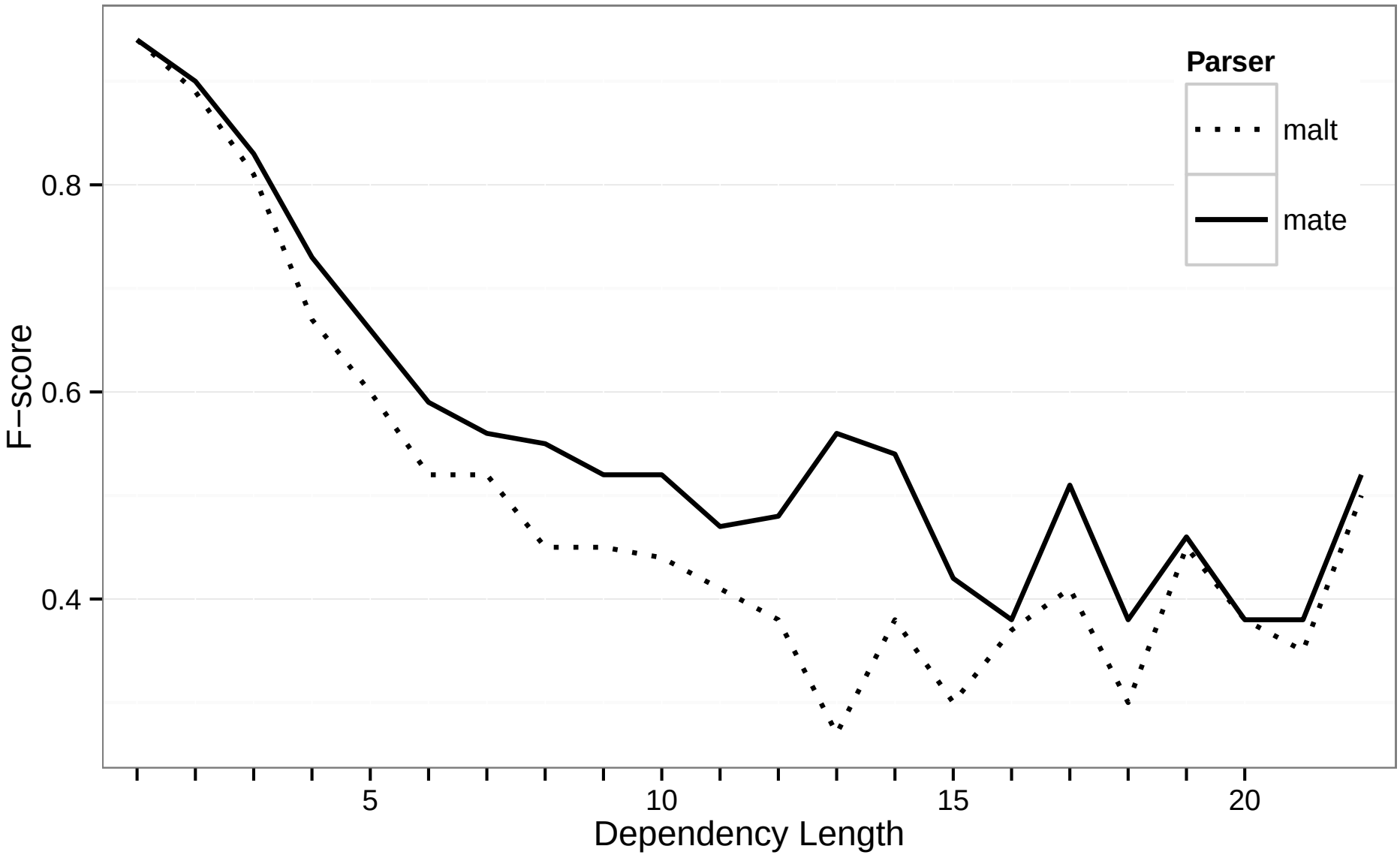
Επίδραση της αυτόματης αναγνώρισης του μέρους του λόγου



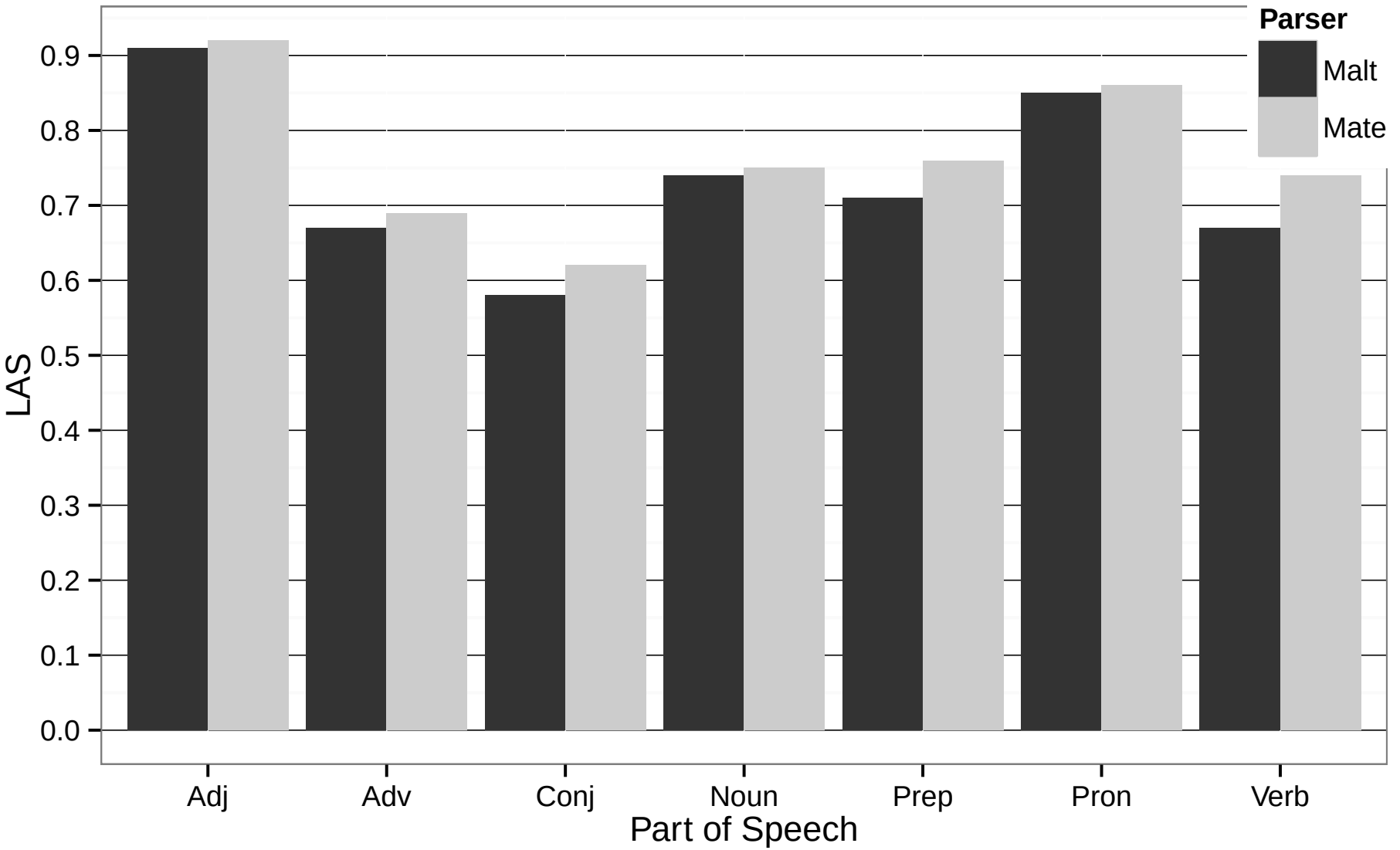
Επίδραση του μήκους της πρότασης



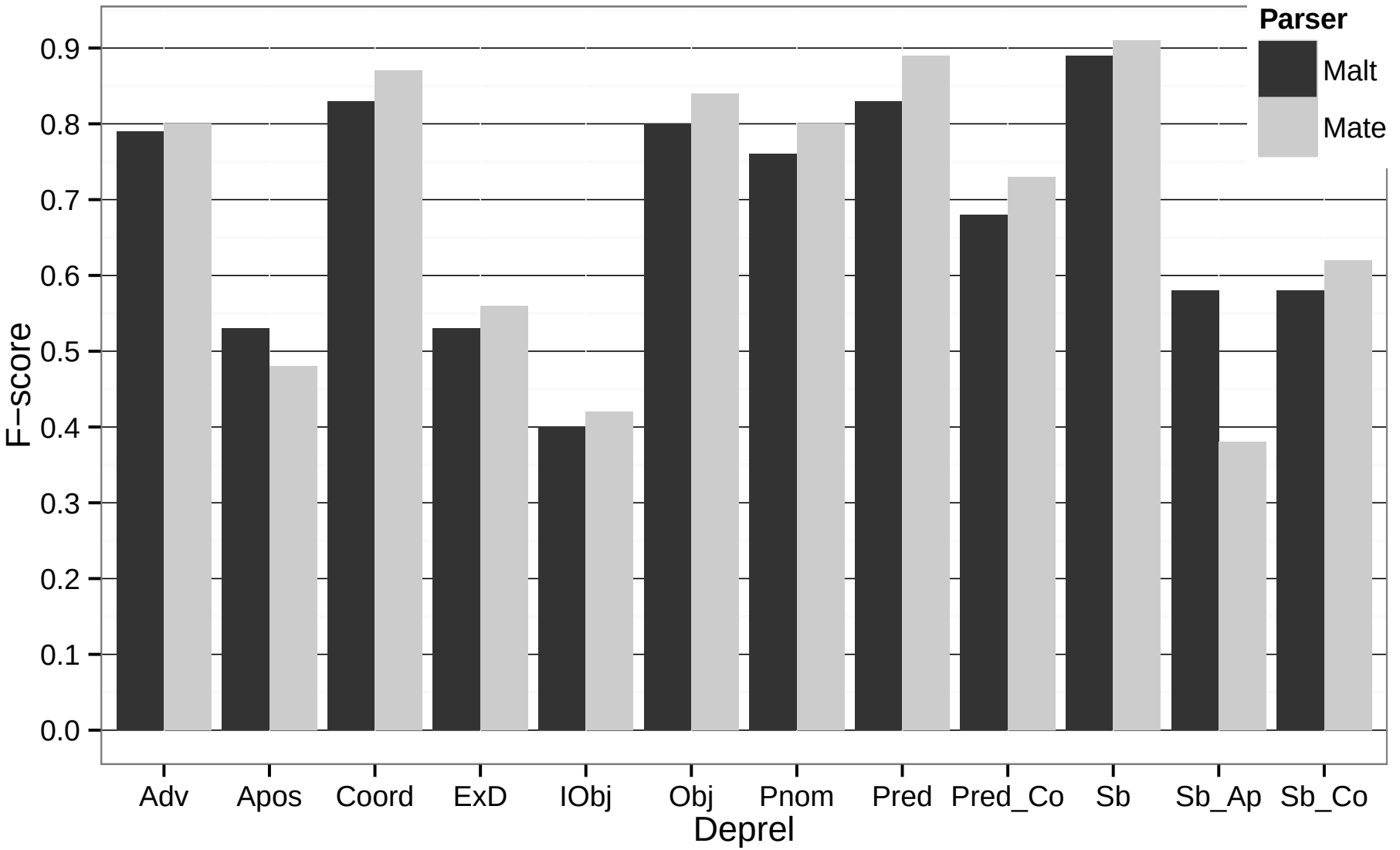
Επίδραση της απόστασης κόμβου/κεφαλής



Επίδραση του μέρους του λόγου



Επίδραση του είδους της σχέσης



Περιεχόμενα

Μείωση μήκους / απλοποίηση κειμένου

- Μείωση μήκους για την παραγωγή υποτίτλων κ.α.
- Απλοποίηση για AMEA

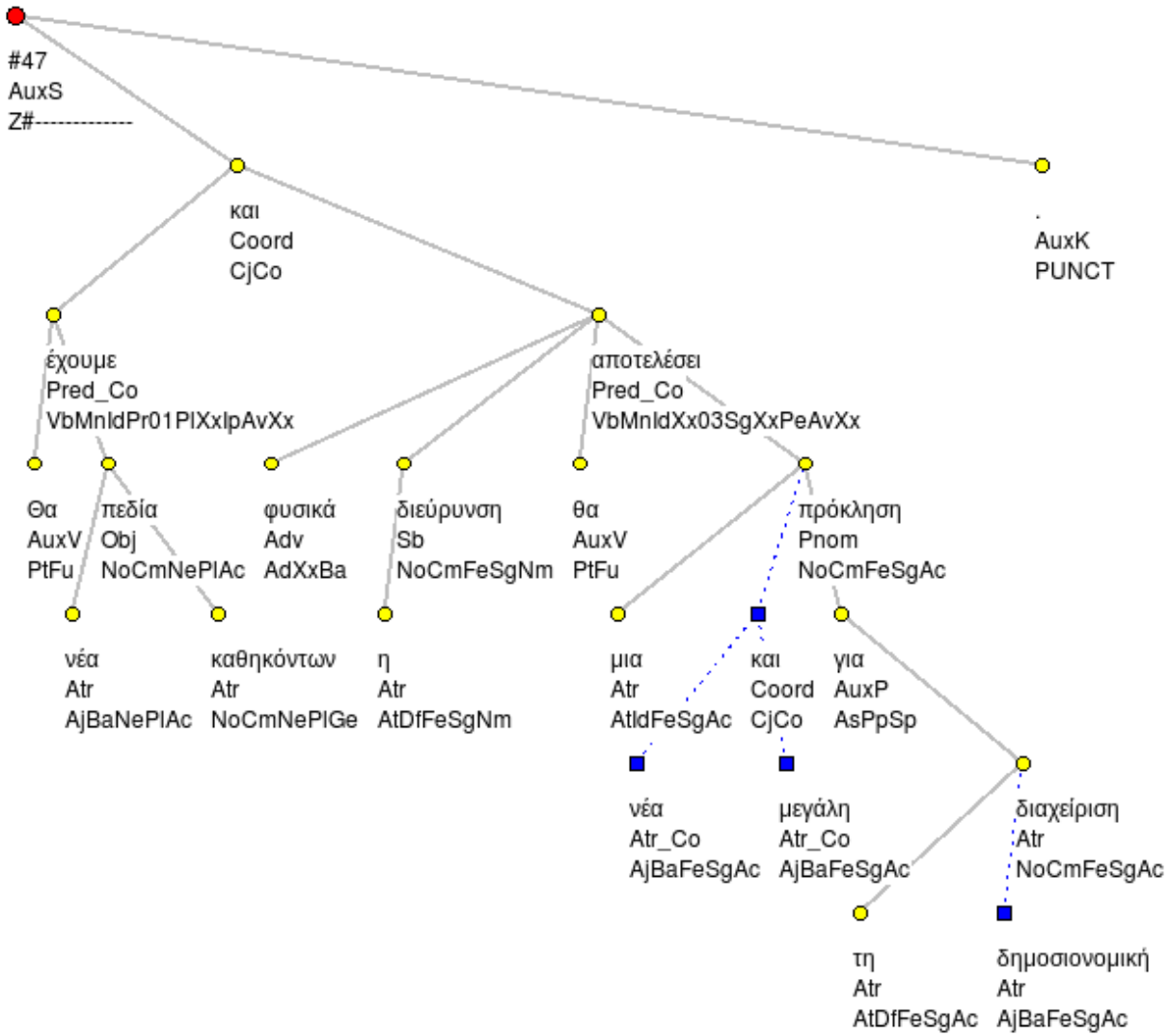
While there are a variety of language problems associated with aphasia, depending on the extent and location of brain damage and the level of pre-aphasia literacy amongst other things, aphasics in general have trouble with long sentences, infrequent words and complicated grammatical constructs including embedded clauses and passive voice [Sid06].

- Υποβοήθηση συστημάτων μηχανικής μετάφρασης και παραγωγής περιλήψεων κ.α.

Κανόνες διαγραφής

- Στόχος η απαλοιφή υποδένδρων στα δένδρα εξάρτησης
  - τα οποία μεταφέρουν δευτερεύουσα σημασιολογική πληροφορία και
  - η απουσία τους δεν αναμένεται να έχει σημαντικές επιπτώσεις στη γραμματική ορθότητα και τη σημασιολογική αποδεκτότητα της πρότασης
- Χρήση κανόνων που διατρέχουν αναδρομικά τους κόμβους του δένδρου, ελέγχοντας αν συγκεκριμένοι μορφοσυντακτικοί περιορισμοί ισχύουν για κάθε κόμβο
- Τα προς διαγραφή υποδένδρα ταξινομούνται με βάση το άθροισμα των συχνοτήτων των λέξεων του υποδένδρου
- Διαγραφή πρώτα των πιο υψίσυχνων (και άρα λιγότερο σημαντικών) συνδυασμών λέξεων

Παράδειγμα εξόδου



Επέκταση: Κανόνες δομικής απλοποίησης

- Μετατροπή αναφορικών προτάσεων σε κύριες

Η επιτροπή καταψήφισε τον νόμο που αφορά την απελευθέρωση του ωραρίου. →
Η επιτροπή καταψήφισε τον νόμο. Ο νόμος αφορά την απελευθέρωση του ωραρίου.

- Μετατροπή των ενεργητικών επιρρηματικών μετοχών σε ρήματα που αποτελούν κεφαλές κύριων προτάσεων

Έχοντας ανάγκη από χρήματα, ο ληστής διέρρηξε την πολυτελή οικία. →
Ο ληστής είχε ανάγκη από χρήματα. Ο ληστής διέρρηξε την πολυτελή οικία.

- Χωρισμός προτάσεων οι οποίες συνδέονται παρατακτικά σε δύο κύριες

Η πρόταση θεωρήθηκε ανεπαρκής και απορρίφθηκε από την επιτροπή →
Η πρόταση θεωρήθηκε ανεπαρκής. Η πρόταση απορρίφθηκε από την επιτροπή.

Συμπεράσματα

- GDT: ένας πολυεπίπεδα επισημειωμένος πόρος για τη NE
- http://gdt.ilsp.gr/
- Χρήση στην εκπαίδευση ενός ΣΑ για εφαρμογές ΕΦΓ
- Ο ΣΑ είναι διαθέσιμος ως web service από το http://nlp.ilsp.gr/ws/ μαζί με άλλα εργαλεία ΕΦΓ με εστίαση στη NE
- Επεκτάσεις:
  - Προσθήκη νέων κειμένων ...
  - ... αλλά και προτάσεων που ικανοποιούν συγκεκριμένα κριτήρια
  - Προσθήκη υλικού από διαφορετικά κειμενικά είδη
  - Επισημείωση και σε άλλα επίπεδα όπως η συναναφορά (επόμενο μάθημα)

Βιβλιογραφία

McDonald, Ryan et al. (2013). Universal Dependency Annotation for Multilingual Parsing. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria,

Nivre, Joakim et al. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In: *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, URL: http://www.aclweb.org/anthology/D/D07/D07-1096.

Prokopidis, Prokopis, Byron Geograntopoulos, and Haris Papageorgiou (2011). A suite of NLP tools for Greek. In: *Proceedings of the 10th International Conference of Greek Linguistics*. Komotini, Greece. URL: http://nlp.ilsp.gr/nlp/ICGL2011_Prokopidis_etal.pdf.

Siddharthan, Advaith (2006). Syntactic Simplification and Text Cohesion. In: *Research on Language and Computation* 4.1. Springer Science,

Zeman, Daniel et al. (2012). HamleDT: To Parse or Not to Parse? In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey,