«Τεχνογλωσσία VIII»
Εξαγωγή πληροφοριών από κείμενα

Σεμινάριο 6: Ανάλυση Πραγματείας

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
στο κεφάλαιο 21 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/

Επίπεδα ανάλυσης

Προ-επεξεργασία	Αναγνώριση λέξεων, προτάσεων, …
Μορφολογική ανάλυση	Πληροφορίες για τις λέξεις, όπως θέμα, κατάληξη, πρόσωπο, αριθμό, γένος, …
Συντακτική ανάλυση	Συντακτική δομή περιόδων, ρόλοι των λέξεων, κλπ.
Σημασιολογική ανάλυση	Αναγνώριση του νοήματος των προτάσεων
Ανάλυση πραγματείας	Αναφορικές εκφράσεις, σχέσεις μεταξύ προτάσεων
Ανάλυση σχεδίου	Σκοποί του χρήστη, σχέδια δράσεως, …

Ανάλυση πραγματείας (1)

- Στην μελέτη της γλώσσας, μερικές από τις πιο ενδιαφέρουσες ερωτήσεις προκύπτουν σε σχέση με τον **τρόπο που χρησιμοποιείται** η γλώσσα
  - Αντί για το ποια είναι τα συστατικά της
- Σημασιολογική ανάλυση:
  - Ασχοληθήκαμε με το πώς οι χρήστες της γλώσσας **ερμηνεύουν** αυτό που άλλοι χρήστες εννοούν

Ανάλυση πραγματείας (2)

- Αν πάμε ένα βήμα παραπέρα, και ασχοληθούμε με το πώς:
  - Βγάζουμε νόημα διαβάζοντας κείμενα
  - Καταλαβαίνουμε τι εννοεί κάποιος άσχετα με το ως το εκφράζει
  - Αναγνωρίζουμε συνδεδεμένες προτάσεις σε σχέση με προτάσεις «ατάκτως ερριμμένες»
  - Συμμετέχουμε σε μια συζήτηση
  - …

→ Τότε εκτελούμε **ανάλυση πραγματείας**

Ανάλυση πραγματείας (3)

- Μελέτη της πραγματικής έννοιας και νοήματος (των προτάσεων/εκφωνημάτων) **πέρα από τη σημασία**
  - Ουσιαστικά αρχίζει εκεί που τερματίζει η σημασιολογική ανάλυση
  - «Η γλώσσα πέρα από την πρόταση ή πέρα από την φράση» (Stubbs, 1983)

Ανάλυση πραγματείας (4)

Η ανάλυση πραγματείας (discourse) ασχολείται με:

- Την μελέτη των **σχέσεων μεταξύ των προτάσεων** (ή τμημάτων τους), ενός μονολόγου ή ενός διαλόγου
  - Του τρόπου με τον οποίο προτάσεις σχηματίζουν μεγαλύτερες μονάδες με νόημα, όπως παραγράφους, μονολόγους, συζητήσεις, κλπ.
  - Του τρόπου με τον οποίο λέξεις/φράσεις/προτάσεις πρέπει να ερμηνευτούν όλες μαζί
- Την μελέτη της **συνεκτικότητας** (coherence) της φυσικής γλώσσας

Παραδείγματα (1)

The Tin Woodman went to the Emerald City to see the Wizard of Oz and ask for a heart. After he asked for it, the Woodman waited for the Wizard's response.

- Τι υποδηλώνουν τα "he", "it";
  - Φαινόμενο: **Αναφορά** (coreference)
  - Εκφράσεις με αναφορά: Αναφορικές εκφράσεις
  - Άρση αμφισημίας: Επίλυση αναφοράς (coreference resolution)

Παραδείγματα (2)

First Union Corp is continuing to wrestle with severe problems. According to industry insiders at Paine Webber, their president, John R. Georgius, is planning to announce his retirement tomorrow.

- Έστω ότι θέλουμε να εξάγουμε μια περίληψη σαν την ακόλουθη:

  First Union President John R. Georgius is planning to announce his retirement tomorrow.

  - Η 2η πρόταση είναι σημαντική
  - Η 1η πρόταση δίνει απλά πληροφορία «υποβάθρου»
- Τέτοιες σχέσεις ονομάζονται **σχέσεις συνεκτικότητας** (coherence relations)

Συνεκτικότητα (coherence)

- Αν από ένα βιβλίο μαζέψουμε μερικές τυχαίες προτάσεις από κάθε κεφάλαιο, έχουμε διήγηση;
  - Όλες οι προτάσεις θα είναι συντακτικά σωστές
  - Όλες οι προτάσεις θα είναι σημασιολογικά σωστές
  - Θα βγαίνει νόημα; (όχι)

«Ο Γιάννης έκρυψε τα κλειδιά του αμαξιού του Νίκου. Ήταν μεθυσμένος.» **(σχέση αιτιολόγησης)**

«Ο Γιάννης έκρυψε τα κλειδιά του αμαξιού του Νίκου. Του αρέσει το σπανάκι.» **(;)**

- Αυτό που λείπει είναι η **συνεκτικότητα**

Συνεκτικότητα (2)

a. John went to his favorite music store to buy a piano.
b. He had frequented the store for many years.
c. He was excited that he could finally buy a piano.
d. He arrived just as the store was closing for the day.

a. John went to his favorite music store to buy a piano.
b. It was a store John had frequented for many years.
c. He was excited that he could finally buy a piano.
d. It was closing just as John arrived.

- Είναι και τα δύο συνεκτικά;
- Είναι και τα δύο το ίδιο συνεκτικά;
  - Μήπως κάποιο είναι περισσότερο συνεκτικό;

Κατάτμηση πραγματείας (1)

- Σε κείμενα, υπάρχουν σημεία που αλλάξει το θέμα
  - Η εύρεση αυτών των σημείων ονομάζεται **κατάτμηση πραγματείας** (discourse segmentation)

Κατάτμηση πραγματείας (2)

- Βοηθάει η μελέτη της συνοχής (cohesion) του κειμένου
  - Η συνοχή (όχι συνεκτικότητα) μεταξύ προτάσεων επιτυγχάνεται μέσω επανάληψης των ίδιων ή σχετικών λέξεων (π.χ. συνώνυμα, υπερώνυμα), κλπ.
    - «Ο Γιάννης καθάρισε την καμινάδα. Δεν ήθελε [ο Γιάννης] άλλα προβλήματα με το τζάκι.»
  - Όσο έχουμε μεγάλη συνοχή, μάλλον δεν αλλάζει το θέμα
- Η συνοχή αφορά τον τρόπο που γλωσσικές μονάδες συνδέονται μεταξύ τους
  - Η συνεκτικότητα αφορά το νόημα των συνδυασμένων μονάδων

Αλγόριθμος TextTiling (Hearst 1997)

- Μια μέθοδος για (γραμμική) κατάτμηση πραγματείας
  - Φυσικά υπάρχουν και άλλες, για γραμμική, ιεραρχική, κλπ. κατάτμηση
  - Η συγκεκριμένη μέθοδος είναι μη επιβλεπόμενη
- Τρία βήματα:
  - Αναγνώριση λέξεων
  - Καθορισμός λεκτικής ομοιότητας
  - Αναγνώριση σημείων αλλαγής θέματος

Αναγνώριση λέξεων

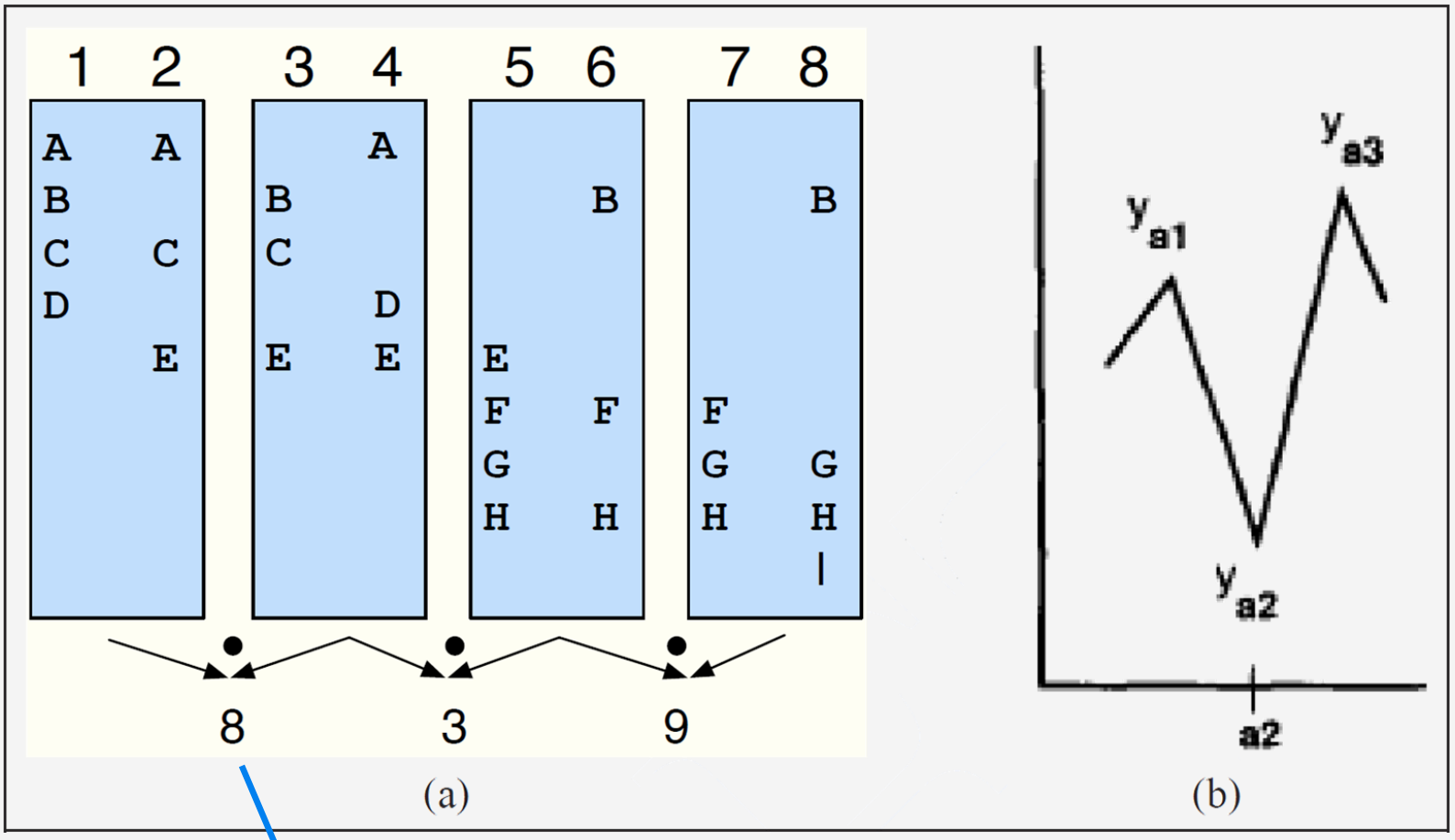
- Κατάτμηση κειμένου στα κενά
- Μετατροπή όλων των λέξεων σε λέξεις με πεζούς χαρακτήρες
- Αφαίρεση συχνών λέξεων (stop-word list)
- Εύρεση θεμάτων (stemming)
- Δημιουργία ψευδό-προτάσεων
  - Αλληλουχίες από 20 θέματα λέξεων

Υπολογισμός λεκτικής ομοιότητας

- Στα σημεία ανάμεσα των ψευδο-προτάσεων, υπολογίζεται η μέση ομοιότητα των λέξεων
  - Εξετάζοντας τις $k = 10$ προηγούμενες και k επόμενες προτάσεις (τα θέματά τους)
  - Δημιουργούνται 2 διανύσματα a, b
    - Που περιέχουν την συχνότητα εμφάνισης κάθε θέματος (από τα N θέματα του κειμένου)
  - Υπολογίζουμε την ομοιότητα μέσω συνημίτονου:

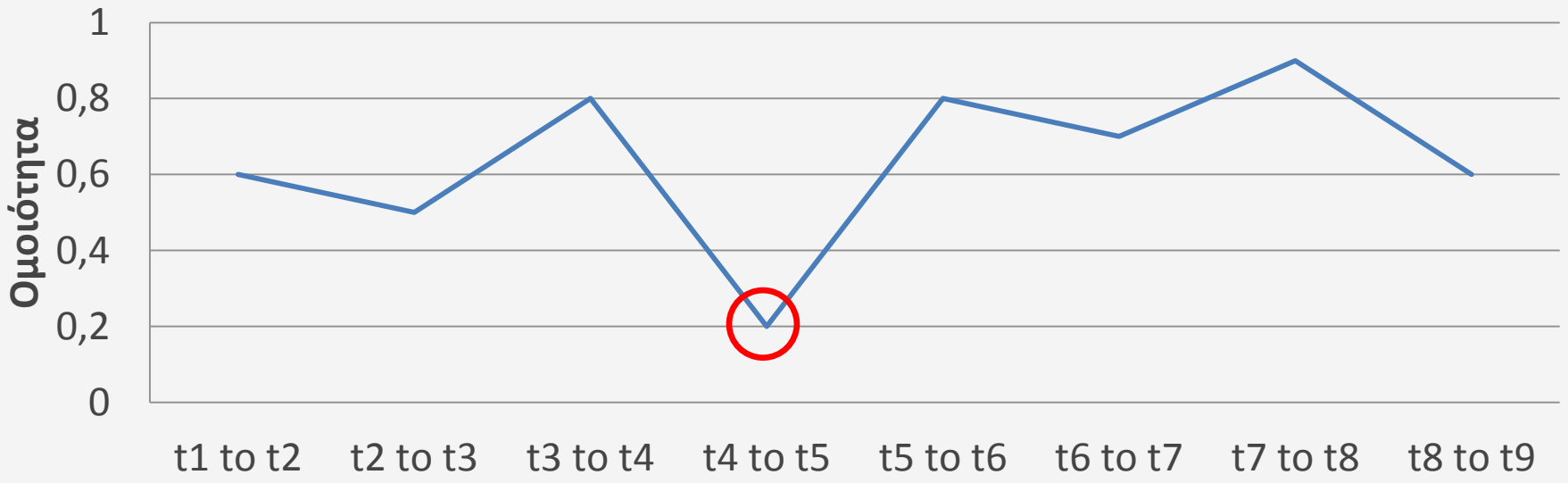
$$sim_{cosine}(\vec{b}, \vec{a}) = \frac{\vec{b} \cdot \vec{a}}{|\vec{b}||\vec{a}|} = \frac{\sum_{i=1}^{N} b_i \times a_i}{\sqrt{\sum_{i=1}^{N} b_i^2} \sqrt{\sum_{i=1}^{N} a_i^2}}$$

Παράδειγμα ($k = 2$)



$$2 * 1 + 1 * 1 + 2 * 1 + 1 * 1 + 2 * 1 = 8$$

Σημεία αλλαγής θέματος



- Εξετάζεται η πτώση $(y_{i-1} - y_i) + (y_{i+1} - y_i)$ από τις δύο γειτονικές κορυφές
- Ευριστικό κριτήριο: υπερβαίνει π.χ. την μέση τιμή + τυπική απόκλιση;
- Εναλλακτικά: ομαδοποίηση (clustering)

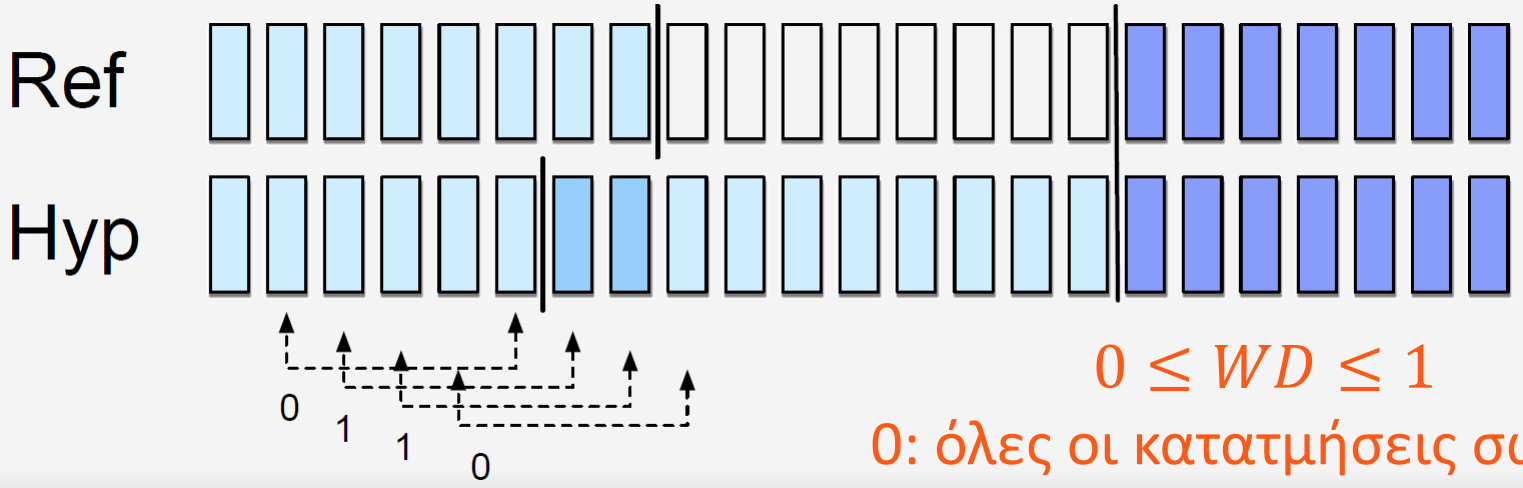
Κατάτμηση με επιβλεπόμενη μάθηση

- Πρόβλημα ταξινόμησης:
  - Αντικείμενα προς κατάταξη: όρια μεταξύ προτάσεων
  - Κατηγορίες: αλλαγή ή όχι θέματος
  - Χαρακτηριστικά:
    - Ομοιότητα μεταξύ προτάσεων
    - Ύπαρξη (ή όχι) συγκεκριμένων λέξεων/φράσεων
      - Όπως χαρακτηριστικές λέξεις/φράσεις (discourse markers, cue phrases)
        - «Καλησπέρα», «Και τώρα ο καιρός», «μετά», «επειδή»…
        - Εύρεση τέτοιων λέξεων/φράσεων με στατιστικές μεθόδους, όπως το πληροφοριακό κέρδος (information gain)

Αξιολόγηση κατάτμησης πραγματείας

- Η ακρίβεια, ανάκληση, f-measure δεν ενδείκνυνται
  - Ίδιο σκορ αν το σφάλμα αφορά μόνο μια πρόταση, με αλγόριθμο που έχει κάνει λάθος αρκετές προτάσεις
- Αξιολόγηση μέσω ολίσθησης παραθύρου k

$$\text{WindowDiff}(ref, hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| \neq 0)$$



$0 \leq WD \leq 1$

0: όλες οι κατατμήσεις σωστές

Συνεκτικότητα (coherence)

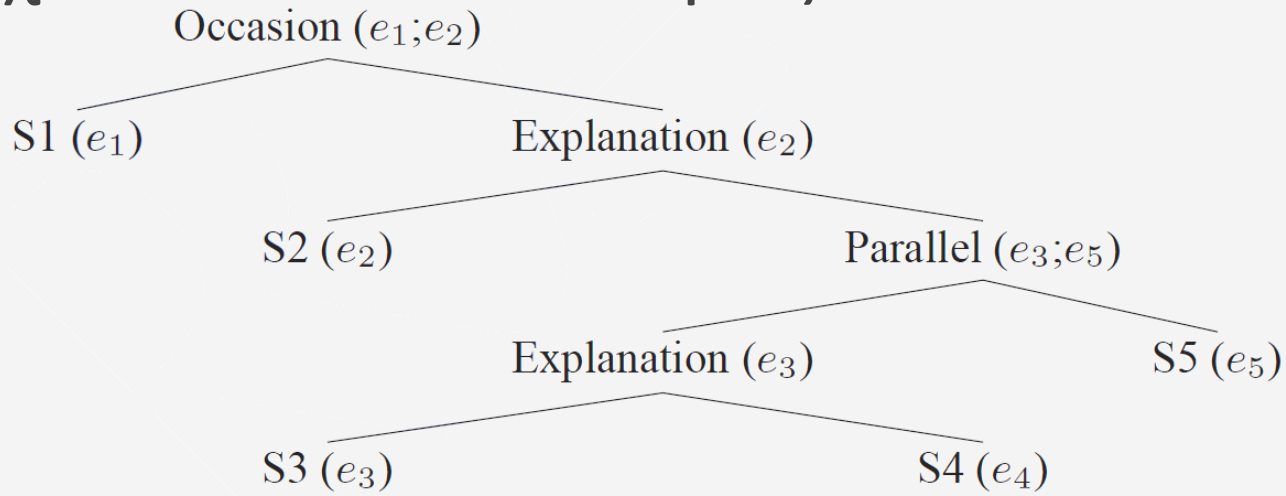
- Οι προτάσεις/φράσεις ενός κειμένου συνδέονται με **σχέσεις συνεκτικότητας** (ή «**ρητορικές σχέσεις**»)
  - Έχουν προταθεί πολλά σύνολα σχέσεων

Σχέσεις συνεκτικότητας

- Οι **σχέσεις του Hobbs** (1979) συμπεριλαμβάνουν:
  - **Αποτέλεσμα** (result): «Άρχισε να βρέχει. Οι αρθρώσεις του Τενεκεδένιου Άνθρωπου σκούριασαν.»
  - **Εξήγηση** (explanation): «Ο Γιάννης έκρυψε τα κλειδιά του αμαξιού του Νίκου. Ήταν μεθυσμένος .»
  - **Παραλληλισμός** (parallel): «Ο Αχυροκεφάλας ήθελε μυαλό. Ο Λαμαρινόκαρδος ήθελε καρδιά.»
  - **Επέκταση** (elaborate): «Η Ντόροθυ ήταν από το Κάνσας. Μεγάλωσε στα λιβάδια.»
  - **Κατάσταση** (occasion): «Η Ντόροθυ σήκωσε το λαδικό. Λάδωσε τις αρθρώσεις του Λαμαρινόκαρδου.»

Δέντρα σχέσεων συνεκτικότητας (1)

- Η συνεκτικότητα ενός κειμένου μπορεί να αναπαρασταθεί από την ιεραρχική δομή μεταξύ των σχέσεων συνεκτικότητας

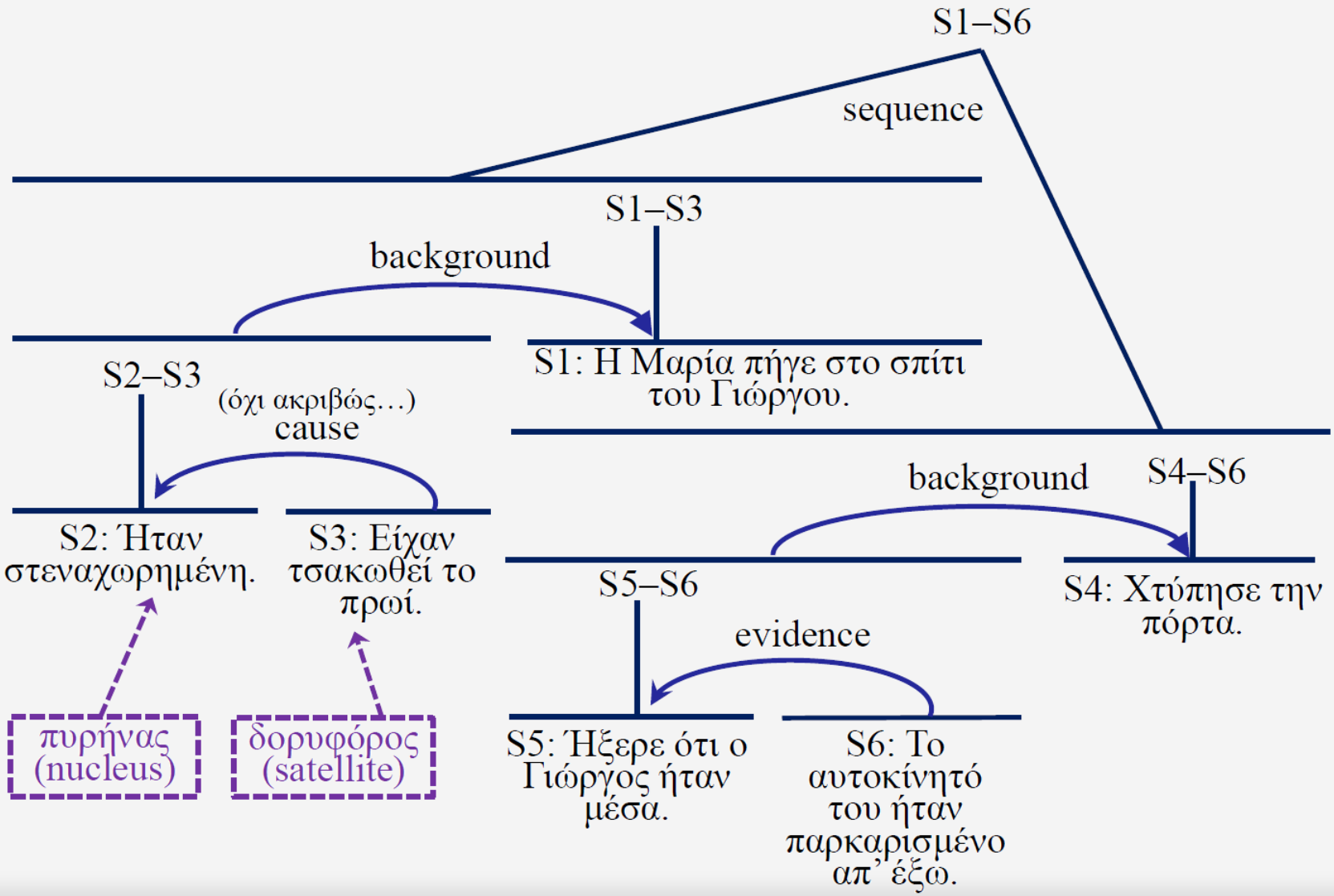


John went to the bank to deposit his paycheck. (S1)
He then took a train to Bill's car dealership. (S2)
He needed to buy a car. (S3)
The company he works for now isn't near any public transportation. (S4)
He also wanted to talk to Bill about their softball league. (S5)

Θεωρία ρητορικής δομής

- Rhetorical Structure Theory (RST, Mann & Thompson 1987)
  - Βασίζεται σε ένα σύνολο από 23 «ρητορικές σχέσεις»
  - Μεταξύ άλλων:
    - Μαρτυρία (evidence): «Ο Γιάννης είναι στο σπίτι του. Το αυτοκίνητό του είναι παρκαρισμένο απ' έξω.»
    - Επέκταση (elaboration): «Ο Γιώργος είναι από την Κύπρο. Μεγάλωσε στη Λευκωσία.»
    - Αντίθεση (contrast): «Ο Γιώργος ήταν χαρούμενος. Η Μαρία ήταν λυπημένη.»
    - Υπόβαθρο (background): «Η Μαρία πήγε στο σπίτι του Γιώργου. Είχαν τσακωθεί το πρωί.»
    - Ακολουθία (sequence): «Η Μαρία πήγε στο σπίτι του Γιώργου. Χτύπησε την πόρτα.»
    - …

Δέντρο ρητορικής δομής



Ανάλυση πραγματείας (1)

- Πώς γίνεται η ανάλυση πραγματείας;
  - Εξαγωγή σχέσεων συνεκτικότητας (ή ρητορικών σχέσεων) μεταξύ δύο προτάσεων (coherence relation assignment)
  - Η εξαγωγή ολόκληρου δέντρου ονομάζεται discourse parsing
- Και οι δύο εργασίες είναι εξαιρετικά δύσκολες
  - Είναι ανοικτά ερευνητικά ζητήματα

Ανάλυση πραγματείας (2)

- Αναγνώριση χαρακτηριστικών λέξεων/φράσεων (discourse markers/cue words)
  - Δρουν σαν «σήματα» μεταβολής της δομής πραγματείας
  - Συχνά αναφέρονται ως «σύνδεσμοι»

«Ο Γιάννης έκρυψε τα κλειδιά του αμαξιού του Νίκου, **επειδή** ήταν μεθυσμένος.»

Ανάλυση πραγματείας (3)

- Κατάτμηση σε τμήματα πραγματείας (discourse segments)
  - Τα οποία δεν είναι πάντα προτάσεις
  - Η συντακτική ανάλυση μπορεί να βοηθήσει στην κατάτμηση

«[Ο Γιάννης έκρυψε τα κλειδιά του αμαξιού του Νίκου], [**επειδή** ήταν μεθυσμένος].»

Ανάλυση πραγματείας (4)

- Αναγνώριση σχέσεων μεταξύ τμημάτων
  - Η αμφισημία των «συνδέσμων» μπορεί να είναι πρόβλημα

«[Ο Γιάννης έκρυψε τα κλειδιά του αμαξιού του Νίκου], [**επειδή** ήταν μεθυσμένος].» (αιτία)

«[Ο Νίκος ήταν μεθυσμένος], [**επειδή** παραπατούσε].» (μαρτυρία)

  - Μπορεί να μην υπάρχουν καν σύνδεσμοι!

«Ο Νίκος ήταν μεθυσμένος. Παραπατούσε.»

Αναφορικές εκφράσεις (1)

The Tin Woodman went to the Emerald City to see the Wizard of Oz and ask for a heart. After he asked for it, the Woodman waited for the Wizard's response.

- Τι υποδηλώνουν τα "he", "it";
  - Φαινόμενο: **Αναφορά** (coreference)
  - Εκφράσεις με αναφορά: Αναφορικές εκφράσεις
  - Άρση αμφισημίας: Επίλυση αναφοράς (coreference resolution)

Αναφορικές εκφράσεις (2)

- Οι αναφορικές εκφράσεις (referring expressions) αναφέρονται κυρίως σε οντότητες του κόσμου

- Αλλά και σε αφηρημένες οντότητες, γεγονότα, κλπ.
  - «Αυτή ήταν μια αισιόδοξη εκτίμηση.»

Είδη αναφορικών εκφράσεων (1)

- Αόριστες αναφορικές φράσεις
  - Αναφέρονται σε συγκεκριμένες οντότητες, κατηγορία οντοτήτων ή γενικευμένο εκπρόσωπο
  - Εισάγουν μια οντότητα στο μοντέλο της πραγματείας
  - «Ο Γιώργος αγόρασε μια τηλεόραση.»
- Οριστικές αναφορικές φράσεις
  - Ανασύρουν μια οντότητα στο μοντέλο της πραγματείας
  - «Ο Γιάννης έδειξε το κινητό στο Νίκο.»
- Κύρια ονόματα
  - Είτε εισάγουν, είτε ανασύρουν

Είδη αναφορικών εκφράσεων (2)

- Αντωνυμίες
  - «Του έδειξε το κινητό. Εκείνος ενθουσιάστηκε.»
- Δεικτικές αντωνυμίες
  - «Θέλω αυτό/εκείνο το κινητό.»

Επίλυση αναφορικών εκφράσεων (1)

- Σε ποια οντότητα (από το περιβάλλον που έχει προηγηθεί) αναφέρεται μια οντότητα;
- Η επίλυση πρέπει να ικανοποιεί κάποια κριτήρια:
  - Συμφωνία γένους, αριθμού και πτώσης
  - Περιορισμούς επιλογής
    - «Πάρκαρε το αυτοκίνητο, αφού το οδηγούσε για ώρες.»
  - Προσφατότητα: συνήθως οι οντότητες που εισήχθησαν τελευταίες σχετίζονται με αναφορές
  - Συντακτικός ρόλος: συμφωνία σε συντακτικό επίπεδο

Επίλυση αναφορικών εκφράσεων (2)

- Η επίλυση πρέπει να ικανοποιεί κάποια κριτήρια:
  - Παραλληλισμός
    - «Η Τασούλα πήγε με την Ελένη για ποτό. Η Βάσω πήγε μαζί της για ψώνια.»
- Διάφορες προσεγγίσεις:
  - Βασισμένες σε κανόνες
    - Ο αλγόριθμός του Hobbs (1978) για αντωνυμίες (Αγγλικά)
    - Θεωρία Επικέντρωσης (Centering Theory)
      - Κάθε στιγμή μόνο μια οντότητα αποτελεί το «κέντρο»
  - Βασισμένες σε μηχανική μάθηση