ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΕΛΟΠΟΝΝΗΣΟΥ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ

6η διάλεξη

Π. ΓΑΚΗΣ

Noam Chomsky

- https://www.youtube.com/watch?v=t-N3In2rLI4&feature=youtu.be&fbclid=IwAR1s0TAt8KjyWTUkmt6y2D5p0odPaFA3QbDDR3b1KrrM3jN3eHgvqbome-4


Ambiguity?
Ambiguity?

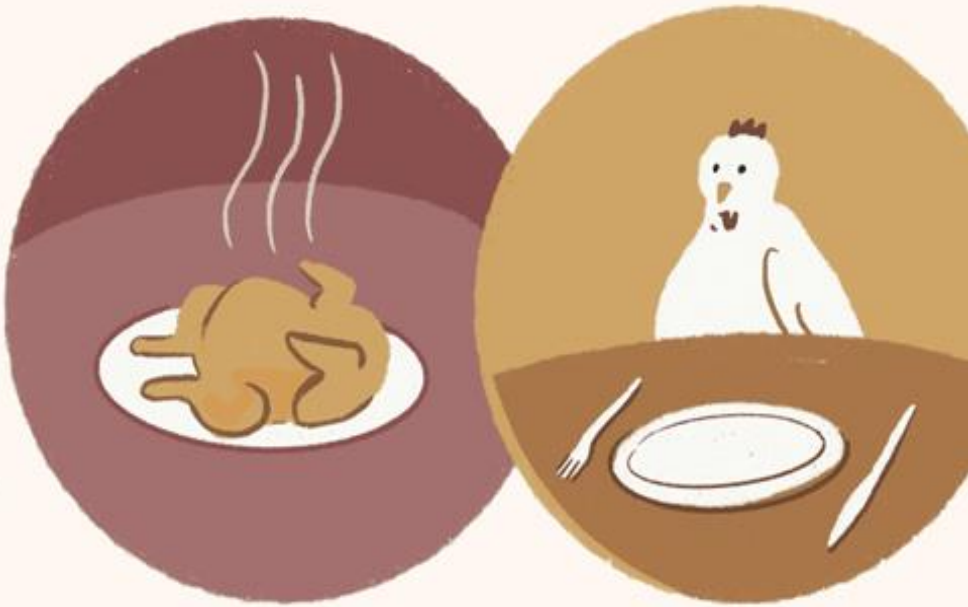
Lexical Ambiguity

The presence of two or more possible meanings within a single word.

"I saw her duck."

Syntactic Ambiguity

The presence of two or more possible meanings within a single sentence or sequence of words.



"The chicken is ready to eat."

ΣΤΟΧΟΙ ΓΛΩΣΣΟΛΟΓΙΚΗΣ ΕΞΕΤΑΣΗΣ

Μέχρι 1980: **έμφαση στη Δομή** της γλώσσας

- Περιγραφές (συγχρονικές/διαχρονικές)

  Φωνητικές/Φωνολογικές
  Μορφολογικές
  Συντακτικές

- Καθολικές αρχές/ Τυπολογία γλωσσών (αναζήτηση της σύγκλισης, των ομοιοτήτων μεταξύ των γλωσσών)

  Language Typology

- Κατάκτηση γλώσσας/Ψυχογλωσσολογία

  Psycholinguistics

- Κοινωνιογλωσσολογία Sociolinguistics

ΓΛΩΣΣΟΛΟΓΙΑ & ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ (NLP)

- Προσέγγιση NLP τα τελευταία 50 χρόνια:

  Υιοθέτηση γλωσσολογικών θεωριών και **ΕΛΕΓΧΟΣ της υπολογιστικής αποτελεσματικότητας** αυτών βάσει εκτεταμένων γλωσσικών δεδομένων με στόχο τη **κατανόηση της φυσικής γλώσσας & την ΑΡΣΗ ΤΗΣ ΑΜΦΙΣΗΜΙΑΣ** (disambiguation)

ΑΞΙΟΠΙΣΤΙΑ & ΕΛΕΓΧΟΣ

Έλεγχος του μοντέλου γλωσσολογικής ανάλυσης ως προς τη

- κάλυψη του φαινομένου
- ανθεκτικότητά του (αντιμετώπιση μη αναμενόμενων δεδομένων)
- πολυπλοκότητα εφαρμογής του ως προς χώρο και χρόνο
- επεκτασιμότητά του
- προσαρμοστικότητά του
- δυνατότητα συντήρησής του

ΜΟΡΦΟΛΟΓΙΑ

Γιατί μορφολογική ανάλυση???

Πιθανές εφαρμογές

1. **Εφαρμογές Φυσικής Γλώσσας (NLP)**

-parsing

-παραγωγή κειμένων

-μηχανική μετάφραση

- λεξικογραφικά εργαλεία & λημματοποίηση

Γιατί μορφολογική ανάλυση???

2. Εφαρμογές Φωνής (Speech applications)

-συστήματα σύνθεσης φωνής

text-to-speech systems

-συστήματα αναγνώρισης φωνής

speech-to-text systems

Γιατί μορφολογική ανάλυση???

3. Εφαρμογές Επεξεργασίας Κειμένου (Word Processing Applications)

-έλεγχος ορθογραφίας – σύνταξης

Spelling checkers – Grammar Checkers

-εισαγωγή κειμένου

Text input (T9 κινητά)

4. Ανάκτηση Εγγράφων (Document Retrieval)

5. Εκπαιδευτικά εργαλεία **διδασκαλίας μορφολογίας** (Ahmad & Rogers 1979; Holman 1988; Klavans & Chodorow 1988)

- **Το είδος της μορφολογικής ανάλυσης που πραγματοποιεί ένας αναλυτής εξαρτάται από την εκάστοτε εφαρμογή που αυτός χρησιμοποιεί**

- ΜΟΡΦΟΛΟΓΙΑ:

- μια δεδομένη λέξη είναι τύπος μιας συγκεκριμένης ρίζας με συγκεκριμένα μορφολογικά χαρακτηριστικά
- σειρά μορφημάτων που εντοπίζονται σε μια μορφολογικά σύνθετη λέξη

ΑΣΑΦΕΙΑ (ΣΗΜΑΣΙΟΛΟΓΙΚΗ)

- Ορισμός
- Τρόποι άρσης σημασιολογικής ασάφειας

Ασάφεια έννοιας λέξεων (Αμφισημία)

- Πολλές λέξεις έχουν αρκετές διαφορετικές έννοιες
  - *Ποντίκι*: τρωκτικό, ηλ.συσκευή, μέρος κρέατος
  - *Γέφυρα*: κατασκευή, προσθετικό οδοντικής
  - *Βιβλιοθήκη*: κτίριο, έπιπλο
- Συχνά η έννοια μιας λέξης γίνεται σαφής από τα συμφραζόμενα μιας πρότασης
  - *Έπιασα στη φάκα ένα ποντίκι.* (τρωκτικό)
  - *Αγόρασα ένα ασύρματο ποντίκι.* (συσκευή)
  - *Αγόρασα ένα κιλό ποντίκι.* (κρέας)
  - *Αγόρασα ένα λευκό ποντίκι.* (τρωκτικό, συσκευή)

Παραδοσιακή Προσέγγιση

- Εισαγωγή συντακτικών-σημασιολογικών περιορισμών στο πώς συνδυάζονται οι λέξεις
  - *Τρώω*: το υποκείμενο πρέπει να είναι ζωντανός οργανισμός και το αντικείμενο κάτι φαγώσιμο
  - *Πράσινος*: μπορεί να προσδιορίζει φυσικά αντικείμενα αλλά όχι αφηρημένες έννοιες
- Οι κανόνες αυτοί καλούνται επιλεκτικοί περιορισμοί (selectional restrictions)

Αποσαφήνιση της έννοιας των λέξεων (Word Sense Disambiguation)

- Η παραδοσιακή προσέγγιση μας επιτρέπει μόνο να ελέγξουμε αν κάτι είναι ή δεν είναι επιτρεπτό
- Στην πραγματικότητα δεν μπορούμε να ορίσουμε πλήρως τις επιτρεπτές σημασιολογικές ιδιότητες μιας έννοιας
  - *Έξυπνος άνθρωπος*
  - *Έξυπνη συσκευή*
  - *Έξυπνες κάρτες*
  - *Έξυπνα πλυντήρια*
- Οι στοχαστικές μέθοδοι μπορούν να βοηθήσουν προς αυτή τη κατεύθυνση βάσει ανάλυσης μεγάλων σωμάτων κειμένων (corpora)

Μοντέλο Μονογράμμου

- Η πιο απλή στοχαστική προσέγγιση είναι να μετρήσουμε πόσες φορές χρησιμοποιείται μία λέξη με την κάθε δυνατή έννοια μέσα σε ένα corpus
  - γέφυρα1 (πρ.οδοντικής): 221 φορές
  - γέφυρα2 (κατασκευή): 4356 φορές
- Αυτές οι μετρήσεις αναφέρονται ως unigrams
- Απαιτούν την ύπαρξη κατάλληλα σχολιασμένου corpus

```
… <wrd sense=like2> like </wrd>   the
<wrd sense=bridge2> bridge </wrd>  of   …
```

N-gram model

- Χρησιμοποιώντας απλά unigrams θα διαλέγαμε ΠΑΝΤΑ την πιο συχνή έννοια για την κάθε λέξη

  (*γέφυρα* → κατασκευή)
- Επομένως πρέπει να λάβουμε υπόψη τα συμφραζόμενα
- Αν s_i είναι η έννοια (sense) της λέξης i
  - Bigrams: $P(s_n \mid s_{n-1})$
  - Trigrams: $P(s_n \mid s_{n-1}, s_{n-2})$

Corpus

- Η ύπαρξη σχολιασμένου corpus δεν είναι πάντα δυνατή
- Πολλοί άνθρωποι διαφωνούν για τις έννοιες συγκεκριμένων λέξεων σε συγκεκριμένα συμφραζόμενα
- Ο ορισμός ενός συνόλου εννοιών δεν είναι ποτέ πλήρης
  - Διαφορετικά επίπεδα εξειδίκευσης

2η προσέγγιση: Λέξεις – Κλειδιά

- Λέξεις που όταν βρίσκονται στα συμφραζόμενα μιας λέξης αποσαφηνίζουν την έννοιά της
- Κάθε λέξη-κλειδί συνδέεται με μία έννοια
  - *δόντι, οδοντίατρος → γέφυρα/πρ.οδοντικής*
  - *πυλώνας, κρεμαστή → γέφυρα/κατασκευή*
- Δεν δίνουν πάντα λύση
  - *Ο οδοντίατρος πήγε στη γέφυρα του Ρίου.*
- Πρέπει κάποιος να ορίσει τις κατάλληλες λέξεις-κλειδιά για κάθε έννοια

3η προσέγγιση: Χρήση λεξικών

- Μπορεί να γίνει χρήση των ορισμών των λεξικών για την εξαγωγή των λέξεων-κλειδιών
- Εναλλακτικά, μπορεί να γίνει σύγκριση των ορισμών διαφορετικών εννοιών δύο λέξεων
  - Επιλέγεται ο συνδυασμός με τη μεγαλύτερη επικάλυψη

Παράδειγμα: Χρήση Λεξικών

"pine cone"

pine 1. kinds of evergreen tree with needle-shaped leaves

2. waste away through sorrow or illness

cone 1. solid body which narrows to a point

2. fruit of certain evergreen trees

Μέγιστη επικάλυψη ορισμών

Λέξεις-κλειδιά: kinds evergreen tree needle-shaped leaves

Λέξεις-κλειδιά: waste sorrow illness

4η προσέγγιση: Χρήση θησαυρών

- Οι θησαυροί συνήθως κατηγοριοποιούν τις λέξεις σε θεματικές κατηγορίες
  - ιατρικός όρος, αθλητικά, κτλ.
- Οι θεματικές κατηγορίες ουσιαστικά είναι οι σημασιολογικές κατηγορίες
- Οι θεματικές κατηγορίες των συμφραζομένων προσδιορίζουν τη θεματική κατηγορία (έννοια) μιας λέξης

ΑΛΛΑ ΕΙΔΗ ΑΣΑΦΕΙΑΣ

ΟΡΙΣΜΟΣ ΚΑΙ ΕΙΔΗ ΑΣΑΦΕΙΑΣ

- 1. **Τυπογραφική ασάφεια**: όταν ένας χαρακτήρας παρουσιάζει περισσότερες από μία λειτουργίες. Χαρακτηριστικό παράδειγμα είναι αυτό της τελείας, η οποία μπορεί να δηλώνει το τέλος μιας πρότασης αλλά μπορεί να εμφανίζεται και σε ένα ακρωνύμιο ή μία συντόμευση χωρίς να σημαίνει απαραίτητα το τέλος της πρότασης.

Π.χ.

Τον ζήτησε χθες ο υπουργός.

Α.Π.Θ.

κ.λπ.

Είδη Ασάφειας

- **2. <u>Μορφολογική ασάφεια</u>:** όταν δύο τύποι του υπολογιστικού λεξικού έχουν την <u>ίδια ορθογραφία</u> και διαφέρουν τουλάχιστον σε ένα από τα παρακάτω χαρακτηριστικά: το λημματικό τύπο, τη γραμματική κατηγορία, την κλίση τους, το γένος ή τον αριθμό, την πτώση.

Κάθε τύπος του υπολογιστικού λεξικού αντιστοιχεί σε έναν κλιτό τύπο μιας λεξικής μονάδας. Για παράδειγμα, σε ένα λεξικό ο τύπος ***<u>τραπέζι</u>*** *(<u>αιτιατική ενικού</u>) είναι διαφορετικός από τον τύπο* ***<u>τραπέζι</u>*** *(<u>κλητική ενικού</u>). Και οι δύο όμως προέρχονται από τον ίδιο λημματικό τύπο*

Είδη Ασάφειας

- 3. **σημασιο-συντακτική ασάφεια:** Αμφίσημη είναι μία γλωσσική έκφραση όταν έχει πολλές σημασίες ή λειτουργίες (Gross,2001)

*Ο καιρός **άλλαξε** (μη μεταβατικό)*

*Ο Γιάννης **άλλαξε** αυτοκίνητο (μεταβατικό, σημασία: αντικαθιστώ)*

*Ο Γιάννης **άλλαξε** το υπνοδωμάτιό του (μεταβατικό, σημασία: δίνω διαφορετική μορφή)*

Είδη Ασάφειας

- **4. <u>Ασάφεια σε κειμενικό επίπεδο</u>**: Μια ακολουθία δύο απλών λεξικών μονάδων σε μία περίσταση επικοινωνίας τη θεωρούμε μία **σύνθετη πολυλεκτική μονάδα**, ενώ σε άλλη περίσταση επικοινωνίας η ίδια ακολουθία μπορεί να αποτελέσει ένα **σύνολο ελεύθερης δομής**.

Π.χ. *Χθες φορούσε μια **μαύρη ζώνη***

*Έχει **μαύρη ζώνη** στο καράτε*

Είδη Ασάφειας

- **5. Λεξική Ασάφεια**: την *ομωνυμία* (π.χ. η αντωνυμία *σου* και το γλυκό *σου* είναι ομώνυμα) και την *πολυσημία* (π.χ. το ρήμα *δίνω* έχει πολλές σημασίες)

- **6. Συντακτική Ασάφεια:** Ένας όρος μιας πρότασης μπορεί να έχει πολλές διαφορετικές συντακτικές δομές

- **7. Σημασιολογική Ασάφεια:** λέξεις με περισσότερες από μία έννοιες στο σημασιολογικό επίπεδο (π.χ. κανάλι: τηλεοπτικός σταθμός || βαθύ αυλάκι || δίοδος επικοινωνίας || δυσκολία

Είδη Μορφολογικής Ασάφειας

- δύο τύποι του υπολογιστικού λεξικού θεωρούνται **μορφολογικά αμφίσημοι** όταν έχουν την ίδια ορθογραφία και διαφέρουν σε **ένα τουλάχιστον** από τα παρακάτω χαρακτηριστικά:

1. τη γραμματική κατηγορία. Για παράδειγμα, ο τύπος **ήπια** αντιστοιχεί στην ονομαστική αιτιατική και κλητική του ενικού του θηλυκού και του πληθυντικού του ουδετέρου του επιθέτου **ήπιος** καθώς και στο πρώτο πρόσωπο του αορίστου του ρήματος **πίνω**

Είδη Μορφολογικής Ασάφειας

2. το λημματικό τύπο. Για παράδειγμα, ο τύπος ***επίπλων*** αντιστοιχεί στη γενική πληθυντικού της λέξης **έπιπλο** και στη γενική πληθυντικού της λέξης **επίπλους**
3. την κλίση τους. Για παράδειγμα, η μονάδα **πάχους** αντιστοιχεί σε δύο τύπους: στη γενική ενικού του λήμματος **πάχος** με ονομαστική πληθυντικού *πάχη* και στη γενική ενικού του λήμματος πάχος με ονομαστική πληθυντικού *πάχη, πάχια και τα πάχητα*

Είδη Μορφολογικής Ασάφειας

4. το γένος. Για παράδειγμα, ο τύπος δικηγόρου αντιστοιχεί στη γενική ενικού του **αρσενικού** και του **θηλυκού** της λέξης *δικηγόρος*
5. τον αριθμό. Για παράδειγμα, ο τύπος γκολ αντιστοιχεί σε όλες τις πτώσεις του ενικού και του πληθυντικού αριθμού της λέξης **γκολ**.
6. την πτώση. Για παράδειγμα, ο τύπος γίγαντα αντιστοιχεί στη γενική και στην αιτιατική ενικού της λέξης **γίγαντας**.

Είδη Ασάφειας

- Γλωσσική ασάφεια προκύπτει και κατά τη μεταφορά του λόγου από το **φωνητικό μέσο στον γραπτό κείμενο**. Για παράδειγμα η λέξη [*λόγια*] (<[**λόγιος**], [**λόγια**]) αναπαριστά δύο διαφορετικές λέξεις που μόνο ο επιτονισμός βοηθά στην αποσαφήνιση, **στοιχείο που χάνεται όμως στον γραπτό λόγο**.

- Ασάφεια είναι δυνατόν να προκύψει και για μια ολόκληρη πρόταση στο ευρύ πλαίσιο της επικοινωνίας (**πραγματολογικό στάδιο**). Και αυτό επειδή η πρόταση είναι ακόμα μια αφηρημένη μονάδα. Έτσι, μια πρόταση όπως: *Θα έρθω το βράδυ* μπορεί, **ανάλογα με το γλωσσικό και εξωγλωσσικό πλαίσιο** στο οποίο θα ενταχθεί και τη **γλωσσική πράξη** στην οποία εγγράφεται, να λειτουργήσει είτε ως απλή δήλωση είτε ως υπόσχεση είτε ως απειλή κ.λπ.

Μορφολογία – Άρση αμφισημίας

- Πώς γίνεται η **άρση** των μορφολογικών **αμφισημιών** που εντοπίζονται σε ηλεκτρονικά σώματα κειμένων (corpus - based approach)?
- με τη δημιουργία **τυποποιημένων γραμματικών κανόνων** και με την εφαρμογή τους στα σώματα κειμένων.
- Η εξάλειψη των αμφισημιών θα επιτρέψει τη **μεγαλύτερη ακρίβεια** στην περιγραφή της γλώσσας, ώστε η επεξεργασία της από ηλεκτρονικές εφαρμογές να είναι αποτελεσματικότερη

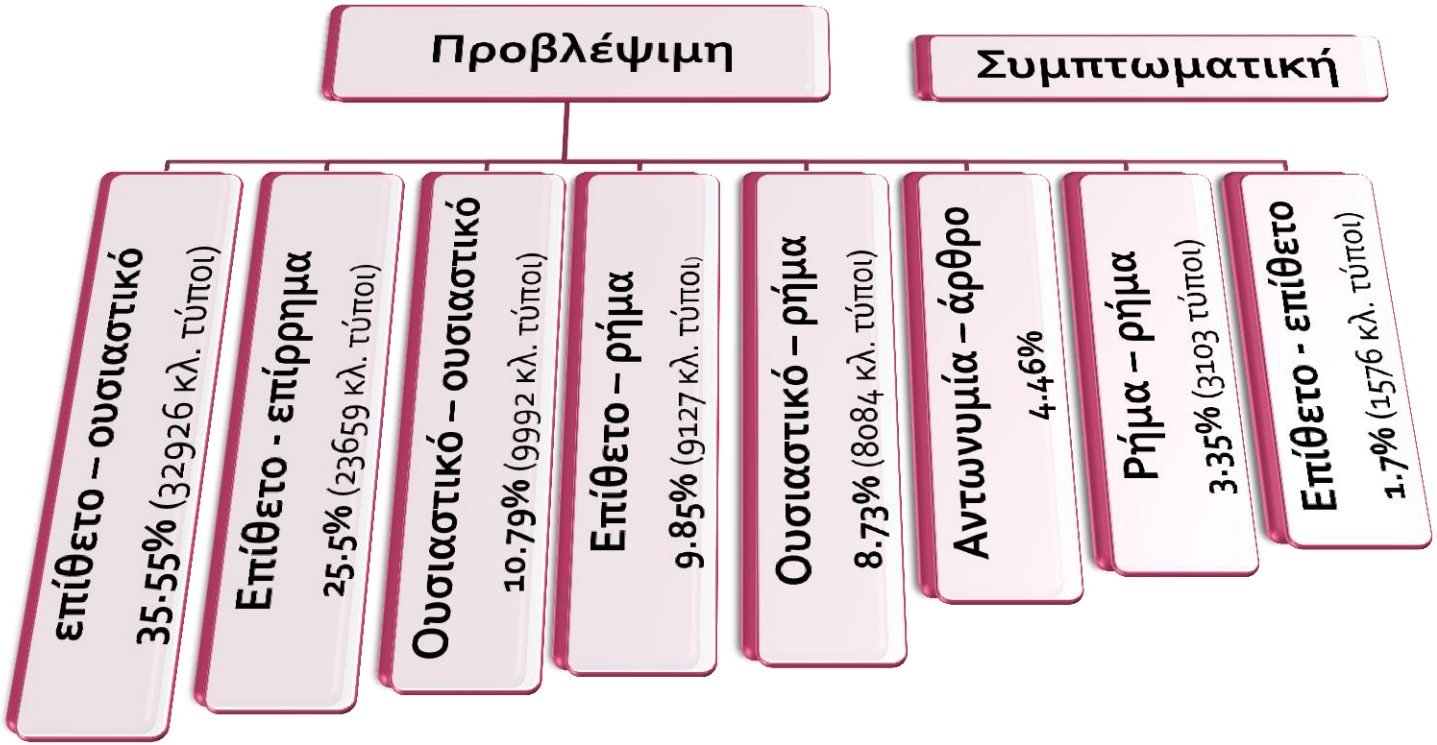
Αίτια λεξικής Ασάφειας

- Η λεξική ασάφεια είναι **προϊόν της πλούσιας μορφολογίας** της ελληνικής γλώσσας.
- Είναι δυνατόν, για παράδειγμα, το λήμμα ενός **ρήματος** να περιλαμβάνει μέχρι και **250-300 τύπους**, συμπεριλαμβανομένων των τύπων των δύο φωνών (ενεργητική, παθητική), του απαρεμφάτου καθώς και των προφορικών τύπων που αυτό έχει στην κλιτική του παραγωγή.
- Αντίστοιχα το λήμμα ενός **επιθέτου** μπορεί να περιλαμβάνει μέχρι και **100 τύπους**, συμπεριλαμβανομένων και των τύπων που παράγονται στα παραθετικά του

Λεξική Ασάφεια

- ο τύπος [*απαντήσεις*] έχει λεξική ασάφεια, γιατί μπορεί να είναι ρήμα (< [*απαντώ*]) ή ουσιαστικό (< [*απάντηση*]).
- Επίσης, ο τύπος [*ματιών*] έχει λεξική ασάφεια γιατί προέρχεται από διαφορετικά λήμματα (< [*μάτι*], [*ματιά*]).
- Επιπλέον, ο τύπος [*κόρη*] παρουσιάζει λεξική ασάφεια, επειδή είναι ασαφής ως προς την πτώση (ονομαστική ή αιτιατική ή κλιτική) μέσα στο ίδιο λήμμα [*κόρη*] και όχι επειδή έχει πολυσημία (κορίτσι || κόρη ματιού).
- Ο ίδιος τύπος δεν έχει λεξική ασάφεια αν, για παράδειγμα, εμφανιστεί στην ονοματική φράση "*η κόρη*", ενώ όμως συνεχίζει να έχει σημασιολογική ασάφεια

Είδη Λεξικής Ασάφειας



Προβλέψιμη Λεξική Ασάφεια

- Η προβλέψιμη ασάφεια περιλαμβάνει δύο κατηγορίες:
- α) η ασάφεια που παρατηρείται **μέσα στο ίδιο το λήμμα**. Ο τύπος [*κρίνω*], για παράδειγμα, μπορεί να είναι οριστική ενεστώτα, υποτακτική ενεστώτα, οριστική εξακολουθητικού μέλλοντα, οριστική συνοπτικού μέλλοντα, καθώς και υποτακτική αορίστου.
- Σε αυτό το επίπεδο ασάφειας ανήκει και η ασάφεια που παρατηρείται μεταξύ της γενικής ενικού του αρσενικού και της γενικής ουδετέρου, καθώς και της γενικής πληθυντικού αρσενικού, θηλυκού και ουδετέρου των αντωνυμιών, επιθέτων και μετοχών.
- Η ύπαρξη κοινών τύπων στο ίδιο λήμμα θεωρείται **εγγενές χαρακτηριστικό του κλιτικού μας συστήματος** και δε σχολιάζεται στις επόμενες παραγράφους, αν και είναι επιβαρυντική για τον υπολογιστή.

Προβλέψιμη Λεξική Ασάφεια

- η ασάφεια που παρατηρείται μεταξύ λεξικών τύπων διαφορετικών λημμάτων με το **ίδιο** ή **διαφορετικό μέρους** του **λόγου**. Υπάρχουν πολλές υποομάδες σε αυτή την κατηγορία (Gakis, Panagiotakopoulos, Sgarbas & Tsalidis, 2013).

1. παροξύτονα ισοσύλλαβα αρσενικά σε **-ας** και στα παροξύτονα ισοσύλλαβα θηλυκά **-α,** που σχηματίζουν κοινούς όλους τους τύπους της κλιτικής τους παραγωγής με μόνη διαφορά τον μορφολογικό χαρακτηρισμό του γένους και της πτώσης. Το μόνο χαρακτηριστικό που παραμένει σταθερό είναι ο αριθμός. Έτσι, έχουμε τους τύπους: {**κεφάλας, κεφάλα, κεφάλες, κεφαλών**} που προέρχονται από τα ουσιαστικά [*κεφάλας*] και [*κεφάλα*] και ανάλογα έχουν τον χαρακτηρισμό (γενική, ενικός, θηλυκό [*κεφάλας*] < [*κεφάλα*]), άλλοτε τον χαρακτηρισμό (ονομαστική, ενικός, αρσενικό [*κεφάλας*] < [*κεφάλας*]).

Προβλέψιμη Λεξική Ασάφεια

2. τα οξύτονα ισοσύλλαβα θηλυκά σε **-ιά** με τα οξύτονα ή παροξύτονα ισοσύλλαβα ουδέτερα σε **-ί/ι** σχηματίζουν κοινή γενική πληθυντικού. ο τύπος [**φασολιών**] προκύπτει από το οξύτονο θηλυκό [φασολιά] και το παροξύτονο ουδέτερο [φασόλι]. Ανάλογη συμπεριφορά έχουν τα: {**σουβλιών** < [σουβλιά, η], [σουβλί, το]}, {**ποδιών** < [ποδιά, η], [πόδι, το]}, {**ματιών** < [ματιά, η], [μάτι, το]} κ.ά. Η ομάδα αυτή των ουσιαστικών αποτελεί το 1,80% της κατηγορίας της λεξικής αμφισημίας ουσιαστικού – ουσιαστικού.

Προβλέψιμη Λεξική Ασάφεια

3. τα οξύτονα ή παροξύτονα ισοσύλλαβα αρσενικά ή θηλυκά σε **–ός/-ος** και τα οξύτονα ή παροξύτονα ισοσύλλαβα θηλυκά σε **-ή** ή σε **–ά** και τα αρσενικά σε **-ας** σχηματίζουν κοινή γενική πληθυντικού: {**αγωγών** < [αγωγός, ο], [αγωγή, η]}, {**γραμματικών** < [γραμματικός, ο], [γραμματική, η]}, {**αρωγών** < [αρωγός, ο], [αρωγή, η]}, {**αυλών** < [αυλός, ο], [αυλή, η]}, {**επωδών** < [επωδός, η], [επωδή, η]}, {**νομών** < [νομός, ο], [νομή, η]}, {**πομπών** < [πομπός, ο], [πομπή, η]}, {**(εθνο)φρουρών** < [(εθνο)φρουρός, ο], [(εθνο)φρουρά, η]}, {**καπνών** < [καπνός, ο], [κάπνα, η]}, {**εμπόρων** < [έμπορος, ο], [έμπορας, ο]}. Τα ουσιαστικά αυτά αποτελούν το 21,88% της ευρύτερης κατηγορίας.

Στοιχεία προβλέψιμης λεξικής ασάφειας

ΛΕΞΙΚΗ ΑΜΦΙΣΗΜΙΑ		
Ουσιαστικό – ρήμα 8.73% (8084 κλ. τύποι)	**Ρήμα – ρήμα** 3.35% (3103 τύποι)	**Ουσιαστικό – ουσιαστικό** 10.79% (9992 κλ. τύποι)
Επίθετο - επίρρημα 25.5% (23659 κλ. τύποι)	**Επίθετο – ρήμα** 9.85% (9127 κλ. τύποι)	**επίθετο – ουσιαστικό** 35.55% (32926 κλ. τύποι)
Επίθετο - επίθετο 1.7% (1576 κλ. τύποι)	**Αντωνυμία – άρθρο** 4.46%	

Συμπτωματική Λεξική Ασάφεια

- Σε αυτή τη μορφή ασάφειας **ταυτίζονται ορθογραφικά λέξεις διαφορετικής ετυμολογικής προέλευσης**.
- Για παράδειγμα, ο τύπος *βάλτε* μπορεί να είναι κλητική ενικού του ουσιαστικού [*βάλτος*] ή προστακτική αορίστου του ρήματος [*βάζω*].
- Επίσης, υπάρχουν τύποι που είναι αδύνατον να αποδοθούν υπολογιστικά με τη σωστή συντακτική ιδιότητα λόγω απουσίας φωνητικής πληροφορίας, όπως ο τύπος *ήπια*, που μπορεί να είναι α΄ ενικό οριστικής αορίστου του ρήματος [*πίνω*] ή ονομαστική, αιτιατική και κλητική πληθυντικού ουδετέρου του επιθέτου [*ήπιος*] ή το επίρρημα [*ήπια*] (αν και η ασάφεια μεταξύ επιθέτου-επιρρήματος είναι προβλέψιμη).

Παραδείγματα Συμπτωματικής Ασάφειας

- τα ουσιαστικά [όρος, το] και [όρος, ο] σχηματίζουν τους κοινούς τύπους [**όρος]** (είτε ονομαστική, αιτιατική, κλητική ενικού ουδετέρου είτε ονομαστική ενικού αρσενικού), [**όρους]** (είτε αιτιατική πληθυντικού αρσενικού είτε γενική ενικού ουδετέρου),
- ασάφεια ουσιαστικού-ρήματος παρατηρείται στους τύπους: {**πελεκάνε** < [πελεκάνος], [πελεκώ]}, {**βάλανε** < [βάλανος], [βάζω]}, {πάταγε < [πάταγος], [πατώ]}, {**βάλτε** < [βάλτος], [βάζω]}, {**έρανε** < [έρανος], [ραίνω]}.
- Συμπτωματική λεξική ασάφεια παρατηρείται και στον τύπο [άσε] που είναι είτε κλητική ενικού του ρήματος [άσος] είτε προστακτική αορίστου του ρήματος [αφήνω]

Παραδείγματα Συμπτωματικής Ασάφειας

- τα ρήματα [δρω] και [εδράζω] σχηματίζουν κοινούς τύπους στον ενικό και στο γ΄ πληθυντικό οριστικής αορίστου: {**έδρασα**, -**ες**, -**ε**, -**αν** < [δρω], [εδράζω]}. Ο τύπος ξέρανε είναι β΄ ενικό προστακτικής αορίστου ή γ΄ ενικό οριστικής αορίστου του ρήματος [ξεραίνω] και γ΄ πληθυντικό παρατατικού του ρήματος [ξέρω]. Ο τύπος (ξανα)δέστε είναι β΄ πληθυντικό προστακτικής αορίστου του ρήματος [(ξανα)βλέπω] και β΄ πληθυντικό προστακτικής αορίστου του ρήματος [(ξανα)δένω]. Ο τύπος [πέστε] είναι προστακτική ενεστώτα ενεργητικής φωνής του ρήματος [λέω] ή προστακτική αορίστου του ρήματος [πέφτω].

Παραδείγματα Συμπτωματικής Ασάφειας

- ο προφορικός τύπος **δω** του επιρρήματος [εδώ] είναι και α΄ ενικό υποτακτικής αορίστου του ρήματος [βλέπω],
- ο τύπος **δικών** της κτητικής αντωνυμίας [δικός] συμπίπτει με τη γενική πληθυντικού του ουσιαστικού [δίκη],
- ο ενικός του θηλυκού της ερωτηματικής αντωνυμίας [πόσος] (**πόση**) συμπίπτει με τον ενικό του θηλυκού ουσιαστικού [πόση],
- ο αδύνατος τύπος **σου** της προσωπικής ή κτητικής αντωνυμίας είναι και ουσιαστικό (ξένη λέξη),
- οι ασθενείς τύποι [**με**] και [**σε**] της προσωπικής αντωνυμίας είναι και προθέσεις,
- ο τύπος **παρά** είναι πρόθεση, συγκριτικός σύνδεσμος και γενική, αιτιατική, κλητική ενικού του ουσιαστικού [παράς].

Στοιχεία συμπτωματικής Ασάφειας

ΣΥΜΠΤΩΜΑΤΙΚΗ ΑΜΦΙΣΗΜΙΑ							
ΜτΛ 1	ΜτΛ 2	Τύποι	Παράδ.	ΜτΛ 1	ΜτΛ 2	Τύποι	Παράδ.
Επιφώνημα	Μετοχή	3	άι	Αντων.	Άρθρο	50	τα
	Επίθετο	12	καλέ		Ουσιαστ.	57	σου
	Επίρρημα	3	ίσα		Επίθετο	39	ίδια
	Ρήμα	3	ορίστε		Ρήμα	8	εμείς
	Ουσιαστ.	104	γιούχα		Επίρρημα	18	κάμποσο
					Σύνδεσμος	3	
Πρόθεση	Αντων.	16	με	Σύνδεσμος	Ουσιαστ.	11	μόλο
	Ουσιαστ	29	συν		Ρήμα	4	όντας
	Adverb	12	υπό		Επίρρημα	33	πριν
					Μετοχή	9	να
					Αντων.	6	όσον
Ουσιαστ.	Ουσιαστ.	5	η				

Άρση Ασάφειας (Disambiguation)

- Τα παραπάνω είδη αμφισημιών **δε δυσχεραίνουν** πάντα την ανθρώπινη επικοινωνία –σε ορισμένες περιπτώσεις δε γίνονται καν αντιληπτές – γιατί γίνεται υποσυνείδητα η επιλογή της σωστής ερμηνείας της λέξης και/ή της πρότασης από τους φυσικούς ομιλητές.
- Για να επιτευχθεί όμως κάτι τέτοιο από τον ηλεκτρονικό υπολογιστή, θα πρέπει οι **κανόνες** που επικαλούμαστε **υποσυνείδητα** κατά την επικοινωνία να **μελετηθούν** και να **καταγραφούν**.
- Οι κανόνες αυτοί θα πρέπει να χαρακτηρίζονται από **συνοχή** (να έχουν ολοκληρωμένη διατύπωση και σαφή δομή), **καθολικότητα** (να ισχύουν σε όλες τις περιπτώσεις των υπό μελέτη φαινομένων) και **συστηματικότητα** (να παράγουν πάντα συγκεκριμένο τύπο αποτελεσμάτων κατά την εφαρμογή τους)

Μορφολογική Ανάλυση

- Για να ξεκινήσει η επεξεργασία ενός κειμένου, το σύστημα πρέπει να έχει πρόσβαση σε μορφολογικές πληροφορίες (Laporte, 2001).
- Στα υπολογιστικά λεξικά, σε κάθε λημματικό τύπο αντιστοιχίζονται όλοι οι πιθανοί κλιτοί τύποι (ακόμη και αν ο λημματικός τύπος ταυτίζεται με τον κλιτό) και κάθε κλιτός τύπος συνοδεύεται από μορφοσημασιολογικές πληροφορίες

Υπολογιστικό λεξικό

- Τα υπολογιστικά λεξικά μπορεί να είναι περισσότερο ή λιγότερα αναλυτικά, με την έννοια ότι περιγράφουν **περισσότερες** ή **λιγότερες** λέξεις/σημασίες των λέξεων, ανάλογα με τη συχνότητα εμφάνισής τους στα σώματα κειμένων.
- Ομοίως, μπορεί να περιλαμβάνουν πολλές ή λίγες πληροφορίες για έναν τύπο (π.χ. μόνο γραμματική κατηγορία ή και πληροφορίες κλίσης). **Όσο πιο αναλυτικά είναι τα λεξικά, τόσο μεγαλώνει ο βαθμός εμφάνισης αμφισημιών** (Courtois, 1996).

- το αποτέλεσμα της εφαρμογής των λεξικών στο εκάστοτε κείμενο, ονομάζεται **λεξική ανάλυση του κειμένου** (lexical analysis ή tagging) και ευθύνεται για την εμφάνιση των αμφισημιών.
- Η λεξική ανάλυση συνίσταται στην αναγνώριση των ελάχιστων μονάδων και στο χαρακτηρισμό αυτών με ετικέτες (tags).
- Η διαδικασία αυτή εξασφαλίζει τη γρήγορη πρόσβαση του συστήματος σε γλωσσολογικά δεδομένα (Laporte, 2001)

- η λεξική ανάλυση ολοκληρώνεται σε δύο φάσεις:

1. Αρχικά, αποδίδονται στις λεξικές μονάδες του κειμένου όλες οι πιθανές ετικέτες με βάση αποκλειστικά τις πληροφορίες των υπολογιστικών λεξικών.
2. Στη συνέχεια, απορρίπτονται κάποιες από αυτές τις ετικέτες με την εφαρμογή γραμματικών κανόνων που περιγράφουν το συγκείμενο. Η εφαρμογή αυτών των κανόνων στα σώματα κειμένων επιτυγχάνει τη μείωση των αμφισημιών

Αναπαράσταση της αμφισημίας στο υπολογιστικό λεξικό

- επίπλων,**έπιπλο**.N:Gns

επίπλων,**επίπλους**.N:Gms

- ήπια,ήπιος.**A**:Nfs:Afs:Vfs:Nnp:Anp:Vnp

  ήπια,πίνω.**V**:J1s

- πάχους[6],πάχος.N:Gns

  πάχους,πάχος.N:Gns

- δικηγόρου,δικηγόρος.N:G**m**s:G**f**s

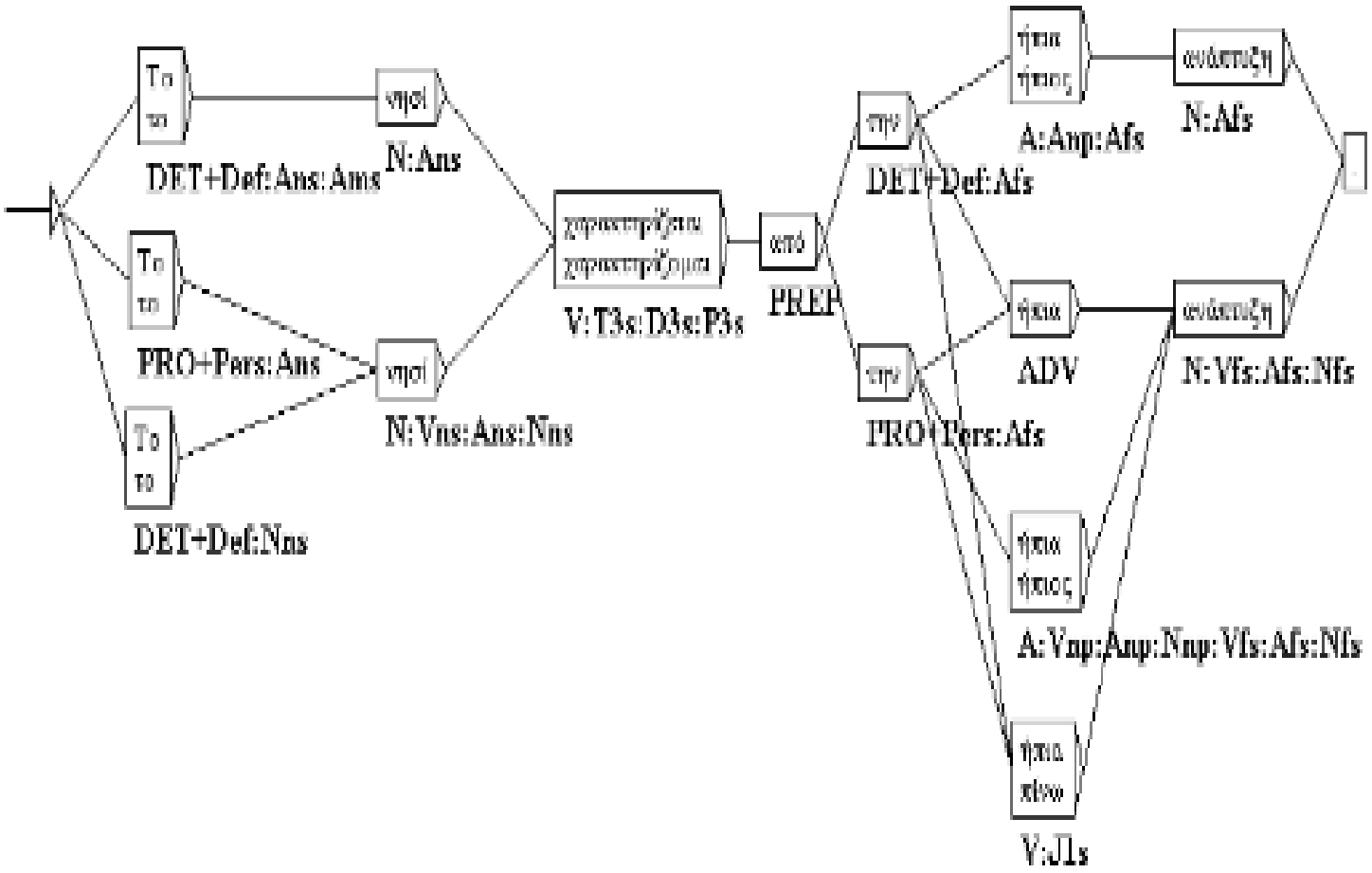
- γκολ,γκολ.N:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp

Ασάφεια και Υπολογιστικά Συστήματα

- Όταν το μέρος του λόγου μιας λέξης είναι **ασαφές**, ο υπολογιστής πρέπει να εξετάσει όλους τους πιθανούς συντακτικούς της ρόλους και, στην περίπτωση που ενεργοποιούνται περισσότεροι από έναν κανόνες, να παράγει όλες τις φραστικές δομές που αυτοί υπαγορεύουν, με την ελπίδα ότι μόνο μία ανάλυση τελικά θα επιτύχει.
- Αυτό όμως προϋποθέτει την επιτυχία στην αναγνώριση γειτονικών δομών, γεγονός που **είναι αμφίβολο**, αν θεωρήσουμε ότι και αυτές μπορεί να περιέχουν ασαφή συστατικά.

Ασάφεια και Υπολογιστικά Συστήματα

- Η ελληνική είναι γλώσσα με πολλές ιδιαιτερότητες, στοιχείο που καθιστά ακόμη πιο δύσκολη και πολύπλοκη την επεξεργασία της από τα συστήματα επεξεργασίας φυσικής γλώσσας. Για παράδειγμα η λέξη [***απαντήσεις*** < **απάντηση**] μπορεί: α) να παίξει το ρόλο της κεφαλής σε μια ονοματική φράση, β) να παίξει το ρόλο κεφαλής σε μια ρηματική φράση [**απαντήσεις** < **απαντώ**]. Επιπλέον ως ουσιαστικό έχει επιπλέον μορφολογική ασάφεια, καθώς μπορεί να είναι ονομαστική ή αιτιατική ή κλητική πληθυντικού.


Το
το
DET+Def:Ans:Ans
νησί
N:Ans
Το
το
PRO+Pers:Ans
νησί
N:Vns:Ans:Nns
Το
το
DET+Def:Nns
χαριστεριέμαι
χαριστεριέμαι
V:T3s:D3s:P3s
από
PREP
την
DET+Def:Afs
ήπια
ήπιος
A:Anp:Afs
ωυάπιξη
N:Afs
την
PRO+Pers:Afs
ήπια
ADV
ωυάπιξη
N:Vfs:Afs:Nfs
ήπια
ήπιος
A:Vnp:Anp:Nnp:Vfs:Afs:Nfs
ήπια
πέφω
V:J1s

<contents>απλά</contents>

<annotations>

<tag name="LEXY" class="class java.lang.String">{απλά,ADV,}{απλός,
ACC+ADJ+NEUT+NOM+PLUR+VOC,CHEM}</tag>

<tag name="TTEXT" class="class java.lang.String">απλά</tag>

<tag name="ORTHO" class="class java.lang.String">NrWrd+WthLtrs+Fvol
Wrd+Style1</tag>

</annotations>

Αποσαφήνιση

- Η **αποσαφήνιση** των χαρακτηριστικών γίνεται από τον **tagger** που αποδίδει τα ορθά μορφολογικά χαρακτηριστικά. Η αποσαφήνιση της λεξικολογικής ασάφειας είναι από τα σημαντικότερα ζητήματα στην επεξεργασία του κειμένου.
- Για παράδειγμα, το να αποφασίσουμε αν το [*απαντήσεις*] είναι ρήμα ή ουσιαστικό μπορεί να επιλυθεί με επισημείωση των μερών του λόγου [**POS**] (***p**art-**o**f-**s**peech tagging*).
- Για παράδειγμα το **context** είναι αυτό που θα καθορίσει εάν ο τύπος [*το*] είναι άρθρο ή αντωνυμία, γνώση απολύτως αναγκαία σε μετέπειτα επίπεδο ανάλυση.
- Το να αποφασίσουμε αν το [*κόρη*] σημαίνει [*κοπέλα*] ή [*κόρη του ματιού*] μπορεί να επιλυθεί με αποσαφήνιση των εννοιών των λέξεων (*word sense disambiguation).*

Ασάφεια

1.Μείωση Αμφισημιών

1.Άρση Αμφισημίας

Μείωση αμφισημιών (ambiguity reduction

- Επομένως , ως μείωση των αμφισημιών (ambiguity reduction) ορίζεται η διαδικασία που εφαρμόζεται στο αποτέλεσμα της λεξικής ανάλυσης (lexical tagging) και που στοχεύει στην **απόρριψη** όσο το δυνατόν **μεγαλύτερου αριθμού λάθος αναλύσεων** με τα απλούστερα και γρηγορότερα δυνατά μέσα (Laporte,2001).
- Το φιλτράρισμα αυτό διευκολύνει τη συντακτική ανάλυση που ακολουθεί, περιορίζοντας τον αριθμό των εναλλακτικών αναγνώσεων μιας πρότασης

Μείωση αμφισημιών (ambiguity reduction

- Π.χ. εξάλειψη των αμφισημιών που αφορούν κυρίως **ουσιαστικά** και **επίθετα** με κοινούς τύπους στις διάφορες **πτώσεις** του ενικού και/ή του πληθυντικού αριθμού καθώς και σε **λέξεις** διαφορετικών γραμματικών κατηγοριών

Παραδείγματα:

Έφυγε χωρίς **τετράδιο**

τετράδιο → N:Nns:Ans:Vns

Παρατήρησε **την** ανοδική πορεία

Μην **την** αντιμετωπίζεις με προκατάληψη

την → DET+Def:Afs \ PRO+Pers:Afs

Υπάρχοντες μηχανισμοί άρσης ασάφειας

- Η υπολογιστική άρση της μορφοσυντακτικής ασάφειας είναι εφικτή μόνο **με την εξέταση των συμφραζόμενων** (context) μιας ασαφούς λέξης. Οι υπολογιστικές μέθοδοι που έχουν αναπτυχθεί με στόχο τη μορφοσυντακτική αποσαφήνιση χωρίζονται γενικά σε **δύο κατηγορίες**:

1. Σύμφωνα με τη γλωσσολογική προσέγγιση, οι ειδικοί κωδικοποιούν **χειρωνακτικά κανόνες** βασισμένους σε γενικεύσεις παραδειγμάτων αποσαφήνισης, τα οποία συνήθως συλλέγονται από σώμα κειμένων μορφοσυντακτικά χαρακτηρισμένων (Ορφανός et al., 1999).
2. Σύμφωνα με την προσέγγιση της μηχανικής εκμάθησης, ένα **στατιστικό μοντέλο** για την επίλυση του γλωσσικού προβλήματος επάγεται **αυτόματα** από σώμα χαρακτηρισμένων κειμένων.

Στατιστικές μέθοδοι

- Οι στατιστικές μέθοδοι βασίζονται στην ανάλυση μεγάλων σωμάτων κειμένων που έχουν σχολιασθεί χειρονακτικά (manually annotated)
- Εξάγονται στατιστικές μετρήσεις
  - *fly*: είναι ουσιαστικό στο 95% των περιπτώσεων που προηγείται άρθρο (*the fly*)
- Αυτές οι μετρήσεις βοηθούν στην ανάλυση νέων (μη-σχολιασμένων) κειμένων
- Χρησιμοποιούμε τη θεωρία πιθανοτήτων για να βρούμε ποια είναι η πιο πιθανή λύση

Υπάρχοντες μηχανισμοί άρσης ασάφειας

- ένας μεγάλος αριθμός διαφορετικών προσεγγίσεων για την άρση της ασάφειας στο ΜτΛ έχει επιχειρηθεί όπως οι προσπάθειες που έγιναν από την ομάδα Δερματά – Κοκκινάκη (Dermatas & Kokkinakis, 1995) και οι οποίες στηρίζονται στη χρήση στοχαστικών συστημάτων και χρησιμοποιούν μοντέλα Hidden Markov (HMM).

  - Η πειραματική διαδικασία γίνεται ξεκινώντας με μέγεθος εκπαίδευσης τις 10Κ λέξεις και αυξανόταν κατά 10Κ κάθε φορά (10.000 λέξεις). Με τη χρήση του μικρού συνόλου ετικετών, το ποσοστό ακρίβειας φτάνει σε αρκετά υψηλά επίπεδα

Υπάρχοντες μηχανισμοί άρσης ασάφειας

- Άλλος τρόπος άρσης της ασάφειας επιχειρήθηκε από **συστήματα που βασίζονται στην εκμάθηση δένδρων απόφασης** την ομάδα των Ορφανού-Χριστοδουλάκη (Orphanos & Christodoulakis, 1999). Επίσης, χρησιμοποιήθηκε και ο αλγόριθμος IGTREE του συστήματος TiMBL, με σκοπό τη σύγκριση της απόδοσής του σε σχέση με τις άλλες παραλλαγές αλγορίθμων εκμάθησης δένδρων απόφασης που χρησιμοποιήθηκαν. Ένα σύνολο κειμένων με μέγεθος 137.765 λέξεων χρησιμοποιήθηκε για την προσέγγιση αυτή

Συστήματα που βασίζονται στην εκμάθηση δένδρων απόφασης

- Το corpus κειμένων αυτό είναι **ποικίλο και ετερόκλητο** και αποτελείται από γραπτά φοιτητών, τμήματα λογοτεχνικών κειμένων, άρθρα από εφημερίδες, τεχνικά, οικονομικά και αθλητικά περιοδικά.
- Έγινε διαχωρισμός των λέξεων και αφέθηκε σε ένα λεξικό η αυτόματη, **πρώτη απόδοση των ετικετών (ΕΠΙΣΗΜΕΙΩΣΗ)**. Το μέγεθος των ετικετών δεν καθορίστηκε αυστηρά, και αφέθηκε στην αρμοδιότητα ενός προγράμματος που συνεργάζονταν με ένα λεξικό να πραγματοποιήσει την αρχική επισημείωση και να επιλέξει το μέγεθος της ετικέτας που θα αποδοθεί στις λέξεις που ανήκουν σε κάθε μέρος του λόγου.
- Ακολούθησε **χειρωνακτική διόρθωση των κειμένων**. Το μοντέλο αυτό, ακολουθώντας τη γλωσσολογική προσέγγιση, επέλυσε την ασάφεια πτώσης, γένους, αριθμού κ.λπ. μέσα από ένα επίπεδο ρηχής συντακτικής ανάλυσης

Παράδειγμα επισημείωσης ΜΤΛ

- Jaguar shares stood at 405 pence before Ford 's initial announcement , but the subsequent takeover frenzy has driven them up.
- Jaguar/NN shares/NNS stood/VBD at/IN 405/CD pence/NN before/IN Ford/NNP 's/POS initial/JJ announcement/NN ,/, but/CC the/DT subsequent/JJ takeover/NN frenzy/NN has/VBZ driven/VBN them/PRP up/RB ./.

Θεωρία πιθανοτήτων

- Αν σε ένα σώμα κειμένων έχουμε
  - 150 εμφανίσεις της λέξης *flies* ως ουσιαστικό
  - 50 εμφανίσεις της λέξης *flies* ως ρήμα
  - $P(\text{category=noun} \mid \text{word=}flies) = 150/200 = 0.75$
- Δημιουργεί πρόβλημα κυρίως στα Αγγλικά
- Οι λέξεις μπορεί να αντιστοιχούν σε περισσότερα από ένα POS tags
  - *The back door* (επίθετο)
  - *On my back* (ουσιαστικό)
  - *Win the voters back* (επίρρημα)
  - *Promised to back the bill* (ρήμα)
- 90% ακρίβεια αν στην κάθε λέξη αποδίδεται πάντα το πιο συχνό tag

Tagger γλωσσολογικά κριτήρια

- Ο tagger (Mnemosyne) που υλοποιήθηκε είναι προσανατολισμένος στην άρση της λεξικής ασάφειας στα νέα ελληνικά.
- Είναι βασισμένος όχι σε στατιστικά στοιχεία αλλά στο ανάλογο γλωσσικό περιβάλλον των λέξεων.
- Η υλοποίηση του tagger έχει προσανατολιστεί στις ανάγκες των κανόνων του parser.
- Αυτό σημαίνει ότι το σύστημα αντιμετωπίζει τον tagger ως υποστηρικτικό αλλά απολύτως απαραίτητο υλικό και δεν αναλύει όλες τις μορφές ασάφειας.
- Υποστηρίζει την πλήρη άρση της λεξικής ασάφειας μόνο με γλωσσολογική πληροφορία και στοχεύει στην ανάδειξή του ως το μοναδικό εργαλείο που διαχειρίζεται τις ασαφείς λέξεις μόνο βάσει του περιβάλλοντός τους.

Tagger Mnemosyne

- Αποτελείται από 70 κανόνες (rules) και η πλειοψηφία των κανόνων αφορά στη άρση της λεξικής ασάφειας μεταξύ άρθρου και αντωνυμίας (34 rules). Από αυτούς τους κανόνες το γλωσσικό περιβάλλον αποδίδει τον μορφολογικό χαρακτηρισμό [άρθρο] ως ΜτΛ σε 20 περιπτώσεις (20 rules) και τον μορφολογικό χαρακτηρισμό [αντωνυμία] ως ΜτΛ σε 14 rules.
- Ο tagger επεκτείνεται και χαρακτηρίζει λέξεις που δεν υπάρχουν στο μορφολογικό λεξικό (άτονες λέξεις, λέξεις με ανορθογραφία) η γνώση των μορφολογικών των οποίων είναι απαραίτητη στο επίπεδο της συντακτικής ανάλυσης

Tagger Mnemosyne

- Οι επιπλέον κατηγορίες άρσης λεξικής ασάφειας είναι οι ακόλουθες:
- Ουσιαστικό και επίρρημα
- Ουσιαστικό και πρόθεση
- Ουσιαστικό και αντωνυμία
- Ουσιαστικό και άρθρο
- Ουσιαστικό και σύνδεσμος
- Ουσιαστικό και ρήμα
- Ουσιαστικό - επιφώνημα
- Ουσιαστικό και μετοχή
- Ρήμα και ουσιαστικό
- Πρόθεση και αντωνυμία
- Επίρρημα και σύνδεσμος
- Επίθετο και ουσιαστικό

Tagger Mnemosyne

- Ο **tagger** του **mnemosyne** δε στοχεύει μόνο στην άρση της ασάφειας ως προς το ΜτΛ αλλά και ως προς το γένος και την πτώση της ασαφούς λέξης. Η απόδοση των σωστών μορφολογικών χαρακτηριστικών **είναι απαραίτητη** σε περιπτώσεις συμφωνίας γένους και αριθμού που εξετάζονται στο επίπεδο της συντακτικής ανάλυσης. Έτσι, αίρει την ασάφεια μεταξύ:

1. Αρσενικού και ουδετέρου
2. Αρσενικού και θηλυκού

- Για την άρση της λεξικής ασάφειας εξετάζονται τόσο οι προηγούμενες λέξεις (σε αριθμό μέχρι και 4 **token**) όσο και/ή οι επόμενες (σε αριθμό μέχρι και 4 **token**)

ΑΡΣΗ ΑΣΑΦΕΙΑΣ (ΒΑΣΙΚΕΣ ΑΡΧΕΣ)

- Αμφισημίες μεταξύ προσδιοριστών και αντωνυμιών μπορούν να εντοπιστούν στους παρακάτω τύπους: ***το, του, της, τον, την, τους, τις, τα*.**

Το πρόγραμμα έχει ως στόχο να προβάλει την πολιτιστική κληρονομιά της κάθε χώρας

To → DET+Def:Nns[8]

Του αποκάλυψε **το** λόγο για τον οποίον ήρθε

To → DET+Def:Ams

Η κοινοτική χρηματοδοτική στήριξη μπορεί να φτάσει μέχρι **το** 50% των συνολικών δαπανών για **το** έργο και δεν μπορεί να υπερβεί το ποσό των 250.000 ECU

To → DET+Def:Ans

Η κυβέρνηση **το** ζήτησε

To → PRO+Pers:Ans

- Ένας τρόπος να διαχωρίσουμε έναν προσδιοριστή από μία αντωνυμία είναι να ορίσουμε ποια μέρη του λόγου έπονται ενός προσδιοριστή. Πράγματι, πριν από ένα επίθετο, ένα ουσιαστικό, έναν αριθμητικό προσδιοριστή, έναν αριθμό ή μία μετοχή, μπορούμε να έχουμε έναν προσδιοριστή, ενώ δεν μπορούμε να έχουμε μία αντωνυμία, παρά μόνο αν είναι κτητική.

Είναι **τα** κορδόνια **των** ωραίων **του** παπουτσιών

Του → PRO+Poss:Gms

- Έτσι, ο γραμματικός κανόνας που θα κατασκευάσαμε διαβάζεται ως εξής:

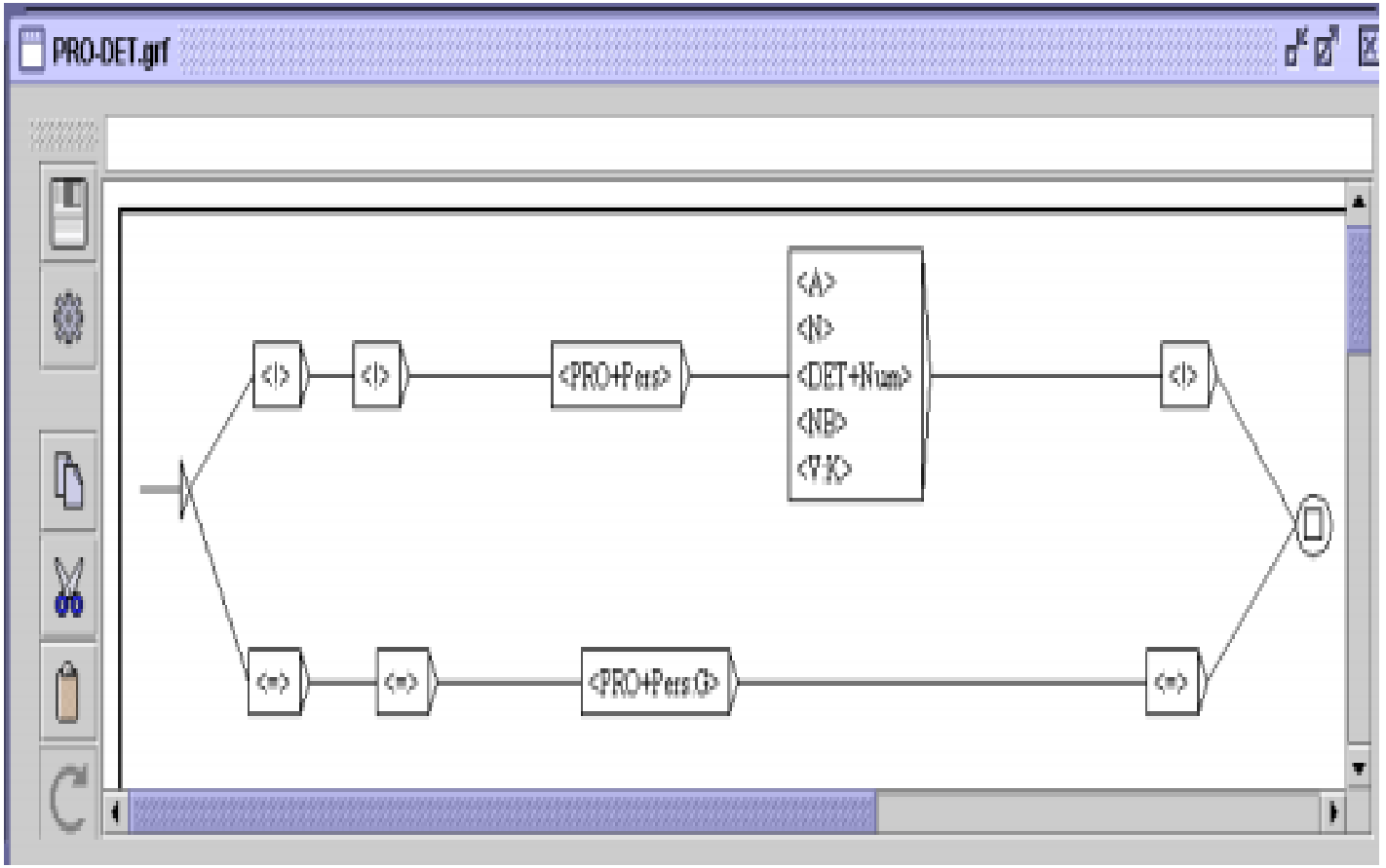
«**Αν συναντήσεις** έναν τύπο που περιγράφεται στα λεξικά και ως αντωνυμία και προηγείται ενός επιθέτου, ενός ουσιαστικού, ενός αριθμητικού προσδιοριστή, ενός αριθμού ή μιας μετοχής (και δεν συμφωνεί σε γένος/αριθμό/πτώση), τότε εφόσον πρόκειται για αντωνυμία, είναι κτητική».

- Με άλλα λόγια, **αν δεν πρόκειται για κτητική** αντωνυμία, τότε η μονάδα για την οποία μιλούμε είναι προσδιοριστής.

Είναι ο οδηγός **του** χτυπημένου αυτοκινήτου

Του → DET+Def:Gns

Γραμματική εξάλειψης αμφισημιών μεταξύ προσδιοριστών και αντωνυμιών



Ασάφειες πτώσεων

- Όσον αφορά τις αμφισημίες που εντοπίζονται στις πτώσεις των ουσιαστικών, ασχολούμαστε μόνο με τις περιπτώσεις όπου η αμφισημία **δεν** οδηγεί σε δύο διαφορετικές καταχωρίσεις. Με άλλα λόγια, **δε μελετώνται μορφολογικές αμφισημίες μεταξύ τύπων** που ανήκουν σε διαφορετική γραμματική κατηγορία, που αντιστοιχούν σε διαφορετικό λημματικό τύπο ή σε διαφορετικό κλιτικό παράδειγμα.

δικηγόρου,δικηγόρος.N:Gms:Gfs

γκολ,γκολ.N:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp

Η ΟΝΟΜΑΣΤΙΚΗ ΠΤΩΣΗ

- Αμφισημίες μεταξύ της ονομαστικής και των άλλων πτώσεων εντοπίζονται σε όλα τα γένη.

γίγαντες	N:Nmp:Amp:Vmp
γυναίκα	N:Nfs:Afs:Vfs
τραπέζι	Nns:Ans:Vns

- Η γραμματική της ονομαστικής που κατασκευάσαμε **περιορίζεται στον ορισμό της ονομαστικής πτώσης των ουσιαστικών από το άρθρο** στην ονομαστική. Εφόσον το αρσενικό και το θηλυκό άρθρο **δεν** είναι ποτέ αμφίσημα, τα χρησιμοποιούμε για να αφαιρέσουμε από τα ουσιαστικά που έπονται τον χαρακτηρισμό της γενικής, της αιτιατικής και της κλητικής πτώσης.

- Ειδικότερα, η εν λόγω γραμματική διαβάζεται από το σύστημα ως εξής: «*Αν συναντήσεις έναν προσδιοριστική, τότε αυτός ο προσδιοριστής βρίσκεται είτε στη γενική είτε στην αιτιατική είτε στην ονομαστική. Στην περίπτωση της ονομαστικής, το ονομαστικό σύνολο που ακολουθείται βρίσκεται και αυτό στην ονομαστική*».

ΓΕΝΙΚΗ ΠΤΩΣΗ

- Για τα περισσότερα ουσιαστικά, η γενική πτώση δε δημιουργεί αμφισημίες με τις άλλες πτώσεις. Τα άκλιτα ουσιαστικά και τα αρσενικά σε –ας αποτελούν τις μόνες εξαιρέσεις. Η γραμματική της γενικής έχει κατασκευαστεί για αυτές τις εξαιρέσεις. Οι περιορισμοί που θέτουμε στη γραμματική αφορούν την ύπαρξη ορισμένων προθέσεων που απαιτούν ουσιαστικό σε γενική (λόγω, μεταξύ, μέσω κ.ό.κ.). Για

Ανέβηκαν με τις σκάλες *λόγω χαλασμένου ασανσέρ*

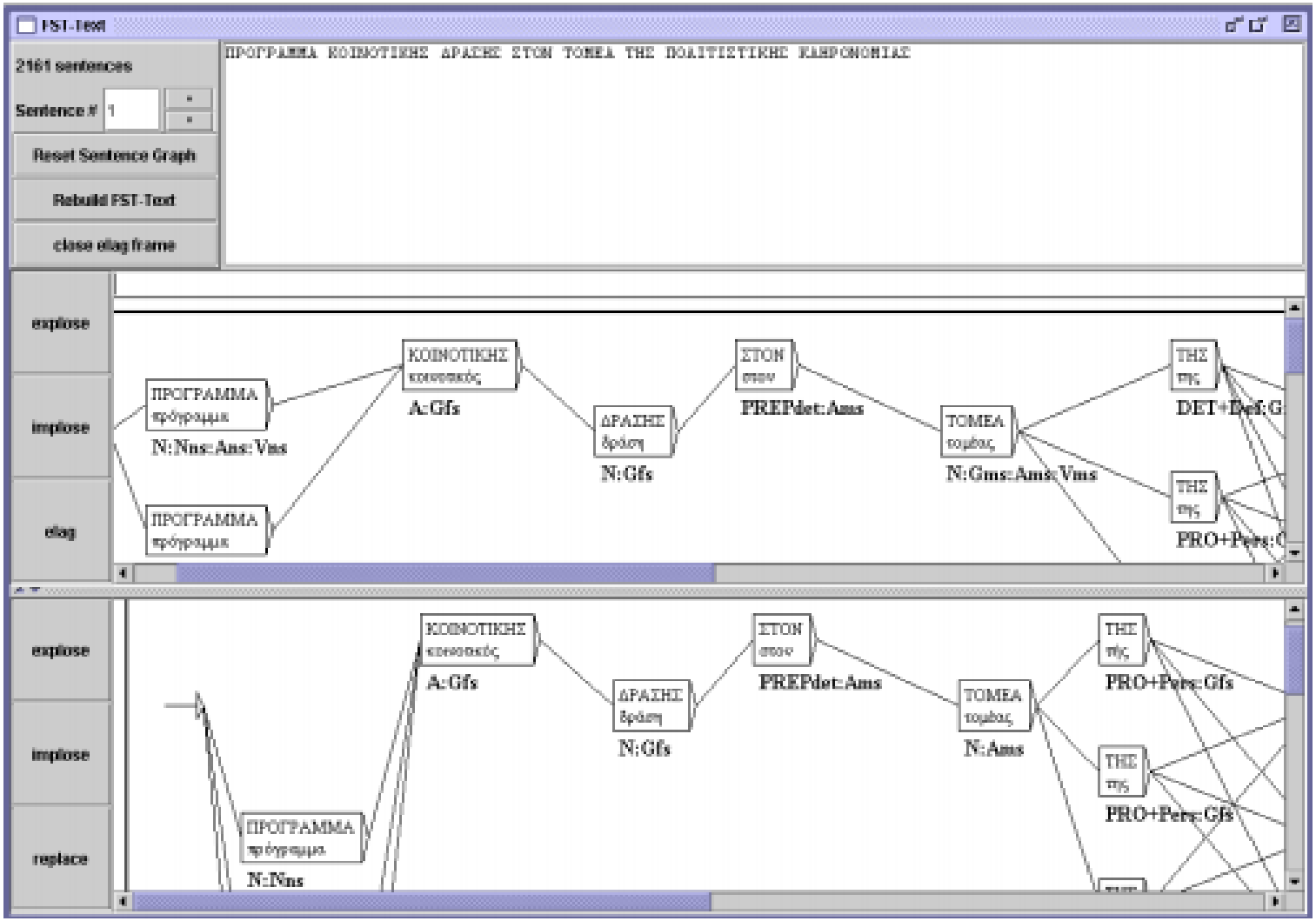
ασανσέρ → N:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp (αμφίσημος τύπος)

ασανσέρ → N:Gns (μη αμφίσημος τύπος)

Η ΑΙΤΙΑΤΙΚΗ ΠΤΩΣΗ

- Προκειμένου να ορίσουμε την αιτιατική πτώση, περιγράψαμε τις ακολουθίες των γραμματικών κατηγοριών που έπονται των προθέσεων. Ειδικότερα, την εν λόγω γραμματική μπορούμε να τη διαβάσουμε ως εξής: «Αν συναντήσεις μία από αυτές τις προθέσεις (**με**, **σε**, **προς** κ.τ.λ.) ή έναν εμπρόθετο προσδιοριστή (**στον**, στην κ.τ.λ.) ακολουθεί ένα ονοματικό σύνολο στην αιτιατική»

Αυτόματο του κειμένου πριν και μετά την γραμματική για την αιτιατική πτώση



```
/* GGC_TAGER_ART_PRONOUN_1*/
1} /
ARULE="No_amb_art_pronoun_1", NULL=CopyTextSpanTagsM(1,[ART]), NULL1=CopyTextSpanTagsN(1,"TTEXT","ORTHO")]  =>
        \
          [LEXY->HasMAmbiguity([ART],[PRON])]
        /
          (
            [TTEXT->Match("σήμερον")],
            [TTEXT->Match("ημέραν")]
          )                                  |
            (
              [TTEXT->Match("φόβον")]        |
              [LEXY->HasMAttrs([CONJ])]
            )
        ;
```

```
/* GGC_TAGER_ART_PRONOUN_2*/   // 2013
{2}
[ARULE="No_amb_art_pronoun_2", NULL=CopyTextSpanTagsN(1,[ART]), NULL1=CopyTextSpanTagsN(1,"TTEXT","ORTHO")]  =>
        [TTEXT->Match(",")]
        \
        [LEXY->HasMAmbiguity([ART],[PRON]), ONTO?=$x:GNC_Agreement(1,[PRON, ART])]
       /
        (
          [LEXY->HasMAttrs([ADJ]), ONTO?=$x:GNC_Agreement(1,[ADJ])]
          [LEXY->HasMAttrs([PART, PASS]), ONTO?=$x:GNC_Agreement(1,[PART])]
        )?,
            [LEXY->HasMAttrs([N]), ONTO?=$x:GNC_Agreement(1,[N])]
       ;
```

```
/* GGC_TAGER_ART_PRONOUN_4*/ // 2013
{1}
[ARULE="No_amb_art_pronoun_4", NULL=CopyTextSpanTagsM(1,[ART]), NULL1=CopyTextSpanTagsN(1,"TTEXT","ORTHO")]  =>
        [LEXY->HasMAttrs([PRON])]
        \
        [LEXY->HasMAmbiguity([ART],[PRON]), ONTO?=$x:GNC_Agreement(1,[ART,PRON])]
       /
        (
          [LEXY->HasMAttrs([ADJ]), ONTO?=$x:GNC_Agreement(1,[ADJ])]
          [LEXY->HasMAttrs([PART, PASS]), ONTO?=$x:GNC_Agreement(1,[PART])]
        )?,
            []{0,1},
            [LEXY->HasMAttrs([N]), ONTO?=$x:GNC_Agreement(1,[N])]
       ;
```

```
    ;
/* GGC_TAGER_N_VERB_4*/
{2} // για τις περιπτώσεις που θέλουμε να διευκρινίσουμε ότι το: απαντήσεις είναι ουσιαστικό ή ρήμα
[ARULE="No_noun_verb_4", NULL=CopyTextSpanTagsM(1,[N]), NULL1=CopyTextSpanTagsN(1,"TTEXT","ORTHO")]  =>
        [LEXY->HasMAttrs([ART])]
        \
        [LEXY->HasMAmbiguity([V],[N])]
        /
        ;
/* GGC_TAGER_N_VERB_5*/
{1} // για τις περιπτώσεις που θέλουμε να διευκρινίσουμε ότι το: απαντήσεις είναι ουσιαστικό ή ρήμα
[ARULE="No_noun_verb_5", NULL=CopyTextSpanTagsM(1,[V]), NULL1=CopyTextSpanTagsN(1,"TTEXT","ORTHO")]  =>
        \
        [LEXY->HasMAmbiguity([V],[INDEC, N])]
        /
        ;
```

```
/* GGC_TAGER_PREP_PRON_1*/  // 2013
{1}
[ARULE="No_amb__PREP_PRON_1", NULL=CopyTextSpanTagsM(1,[PREP]), NULL1=CopyTextSpanTagsN(1,"TTEXT","ORTHO")]  =>
        \
          [LEXY->HasMAmbiguity([PREP],[PRON])]
        /
          [LEXY->HasMAttrs([ACC])]  |
          [LEXY->HasMAttrs([GEN])]
        ;

/* GGC_TAGER_ADJ_N_3*/
{1} // για τη φράση: πιο πρακτικό επίπεδο (βλέπει το πρακτικό ως ουσιαστικό και ως επίθετο
[ARULE="No_adj_n_3", NULL=CopyTextSpanTagsM(1,[ADJ]), NULL1=CopyTextSpanTagsN(1,"TTEXT","ORTHO")]  =>
        \
          [LEXY->HasMAmbiguity([ADJ],[N]), ONTO?=$x:GNC_Agreement(1,[ADJ, N])]
        /
          [LEXY->HasMAttrs([N]), ONTO?=$x:GNC_Agreement(1,[N])]
        ;
```

ΣΥΜΠΕΡΑΣΜΑΤΑ

- Οι γραμματικές που παρουσιάσαμε έχουν επαληθευτεί σε μεγάλα ηλεκτρονικά σώματα κειμένων. Ωστόσο, οι αμφισημίες που μελετήσαμε αποτελούν ένα μικρό μέρος του συνόλου των αμφισημιών που παρουσιάζει η Νέα Ελληνική.
- Οι υπόλοιπες αμφισημίες θα πρέπει να εξαλειφθούν είτε σε μορφολογικό επίπεδο, με την κατασκευή περισσότερων γραμματικών που να εφαρμόζονται σε συνδυασμό με αυτές που παρουσιάσαμε παραπάνω, είτε σε συντακτικό επίπεδο, με τη χρήση του λεξικού-γραμματικής

ΣΥΜΠΕΡΑΣΜΑΤΑ

- Οι γραμματικές εξάλειψης αμφισημιών μπορούν να χρησιμοποιηθούν σε πολλές εφαρμογές, όπως η μετατροπή γραπτού σε προφορικό λόγο, ο ορθογράφος, η μηχανική μετάφραση, η αυτόματη περίληψη, οι οποίες δε λειτουργούν σε ικανοποιητικό επίπεδο.
- Οι εφαρμογές αυτές, θα πρέπει να επαληθεύονται τακτικά σε μεγάλα σώματα κειμένων και με ενημερωμένα λεξικά.


Καλή Ανάσταση
Το άγιο Φως της Ανάστασης ,
ας φωτίσει τις ψυχές
και τις ζωές όλων μας.