ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΕΛΟΠΟΝΝΗΣΟΥ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ

1η διάλεξη

Π. ΓΑΚΗΣ

ΣΤΟΧΟΙ ΓΛΩΣΣΟΛΟΓΙΚΗΣ ΕΞΕΤΑΣΗΣ

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

- Περιγραφές (συγχρονικές/διαχρονικές)

Φωνητικές/Φωνολογικές

Μορφολογικές

Συντακτικές

- Καθολικές αρχές/ Τυπολογία γλωσσών
  Language Typology
- Κατάκτηση γλώσσας/Ψυχογλωσσολογία
  Psycholinguistics
- Κοινωνιογλωσσολογία Sociolinguistics

Δευτερεύουσας σημασίας:

- Σημασιολογία Semantics
- Πραγματολογία Pragmatics
- Ανάλυση λόγου Discourse Analysis
- **ΑΜΦΙΣΗΜΙΑ ΑΠΟΔΕΚΤΗ**

ΓΛΩΣΣΟΛΟΓΙΑ & ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ (NLP)

- Προσέγγιση NLP τα τελευταία χρόνια:

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

ΜΕΘΟΔΟΙ NLP & ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΓΛΩΣΣΟΛΟΓΙΑΣ

- Βασισμένες στη γνώση rule/knowledge-based
- Στατιστικές data-driven

Μοντέλα και αλγόριθμοι από:

-Επιστήμη των Η/Υ

-Γλωσσολογία

-Τεχνητή Νοημοσύνη

-Λογική

-Μαθηματικά

-Ψυχολογία

-Φιλοσοφία

-Γνωσιακή Επιστήμη (cognitive science) είναι το επιστημονικό πεδίο που ασχολείται με τη μελέτη του νου)

ΒΑΣΙΚΑ ΕΡΓΑΛΕΙΑ NLP για ανάλυση φωνητική, μορφολογική & συντακτική

Τυπικά συστήματα κανόνων : δηλωτικά

- Κανονικές Γραμματικές (Regular Grammars) & Κανονικές Σχέσεις (Regular Relations)
- Αλγεβρικές Γραμματικές (Context-Free Grammars)
- Γραμματικές Επαυξημένες με Χαρακτηριστικά (Feature-Augmented Grammars)

  + παραλλαγές αυτών με στοιχεία πιθανοτήτων

Υπολογιστική Γλωσσολογία (ΥΓ): Γιατί;

Τί είναι η ΥΓ;

- Η σπουδή υπολογιστικών συστημάτων για την κατανόηση και παραγωγή φυσικών γλωσσών (ένα αντικείμενο που παντρεύει ανθρωπιστικές και θετικές επιστήμες)

Γιατί να ασχοληθεί κανείς με την ΥΓ;

- έχει άμεση σχέση με τη διευκόλυνση της επαφής ανθρώπου-μηχανής

Τι απαιτείται για τις σπουδές σε ΥΓ;

- Για εμάς, το υπόβαθρο στη γλωσσολογία είναι αρκετό.

ΙΣΤΟΡΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

- Οι πρώτες έρευνες, που θα μπορούσαμε να θεωρήσουμε ότι ανήκουν στο πεδίο της Υπολογιστικής Γλωσσολογίας, ξεκίνησαν αμέσως μετά τον Β΄ Παγκόσμιο Πόλεμο σε τομείς όπως τα ***αυτόματα*** (automata) και τα ***στατιστικά ή πληροφοριακά μοντέλα*** (probabilistic or information theoretic models). Από τη δουλειά αυτή προέκυψε η θεωρία των τυπικών γλωσσών όπου σημαντική θέση κατέχει η λεγόμενη «Ιεραρχία του Chomsky» (Chomsky, 1956).

ΙΣΤΟΡΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

- Χρήση Η/Υ στη **λογοτεχνική** ανάλυση

  Literary & Linguistic Computing
- **Μηχανική Μετάφραση (ΜΜ)**

  Machine Translation(MT)
- 17ο αιώνα: πρόταση Descartes & Leibniz: Ανάγκη δημιουργίας λεξικών βασισμένων σε καθολικούς αριθμητικούς κώδικες

ΙΣΤΟΡΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Η έρευνα περιόδου 1956-1966 είναι σημαντικότατη, όχι μόνο για τη ΜΜ, μα κυρίως για την Υπολογιστική Γλωσσολογία και την Τεχνητή Νοημοσύνη (**ανάπτυξη αυτοματοποιημένων λεξικών και τεχνικών συντακτικής ανάλυσης**)

Σημαντική συμβολή στη Γλωσσολογία

Συνέδρια - Περιοδικά

- Τα σημαντικότερα συνέδρια του χώρου είναι τα: ACL, EACL, NAACL (North American Association for Computational Linguistics), COLING (International Conference on Computational Linguistics), LREC (Language Resources Evaluation Conference), LACL (Logical Aspects of Computational Linguistics) και EMNLP (Empirical Methods on Natural Language Processing).
- Τα σημαντικότερα περιοδικά του χώρου είναι τα Computational Linguistics και Natural Language Engineering.

Υπολογιστική Γλωσσολογία (ΥΓ): Πρακτικά..τί παράγεται;

ΥΓ και δεδομένα (Korpuslinguistik)

- Το 2003 η ετήσια παραγωγή έφτανε τα 8 terrabytes (8000 Gigabytes ή 8000 φορτηγάκια γεμάτα βιβλία).

ΥΓ και δεδομένα (Korpuslinguistik)

- ένας άνθρωπος θα χρειαζόταν 5 χρόνια για να διαβάσει ό,τι επιστημονικό παράγεται σε 24 ώρες

- *Ο μόνος τρόπος για την αντιμετώπιση της έκρηξης πληροφοριών και παρακολούθηση της εξέλιξης* ***είναι η αξιοποίηση υπολογιστικών συστημάτων*** *για το χειρισμό τεράστιων ποσών ηλεκτρονικής πληροφορίας*

ΥΓ - δεδομένα (κοινωνική διάσταση)

- Τι κανονικότητες εξάγονται μέσα από τα διάφορα είδη δημοσίου λόγου;
- Πώς επιδρά το εξωγλωσσικό περιβάλλον στη διαμόρφωση κειμενικών ειδών;
- Τι συμπεράσματα μπορούν να εξαχθούν σχετικά με τις γλωσσικές συνήθειες – συμπεριφορές των μελών διαφόρων κοινοτήτων;

ΥΓ - δεδομένα (κοινωνική διάσταση)

Η συστηματική και αυτοματοποιημένη αξιοποίηση επισημειωμένων κειμένων με υπολογιστικά μέσα βοηθά στην κατεύθυνση για καλύτερη

α) πρόβλεψη και αποφυγή «παγίδων» και

β) ενδοσκόπηση στα κίνητρα και ιδεολογίες των συνομιλούντων ή/και συγγραφέων.

Υπολογιστική Γλωσσολογία (ΥΓ): Ακόμα πιο πρακτικά … τι υπάρχει ήδη και τι μέλλει να γίνει.

- Καναδικό υπολογιστικό σύστημα δέχεται καθημερινά καιρικά δεδομένα και αναπαράγει καιρικές αναφορές σε γαλλόφωνο και αγγλόφωνο κοινό
- Επισκέπτες του Cambridge της Μασαχουσέτης μπορούν να ρωτήσουν ένα υπολογιστή για πιθανά εστιατόρια και να διεξαγάγουν διάλογο σχετικά με το μενού και τις τιμές του.
- Υπολογιστικό σύστημα δέχεται εκατοντάδες εργασίες φοιτητών και τις διορθώνει αυτόματα με τρόπο που να μη διαφοροποιείται από ανθρώπινο διορθωτή
- Υπολογιστικό σύστημα δέχεται video μαγνητοσκοπημένων αθλητικών μεταδόσεων ή/και άλλων δραστηριοτήτων και απομονώνει σκηνές που του ζητούνται προφορικά από το χρήστη
- Υπολογιστικό σύστημα βοηθά άτομα με ειδικές ανάγκες για την παραγωγή/εκφορά λόγου.

ΥΠΟΛΟΓΙΣΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ

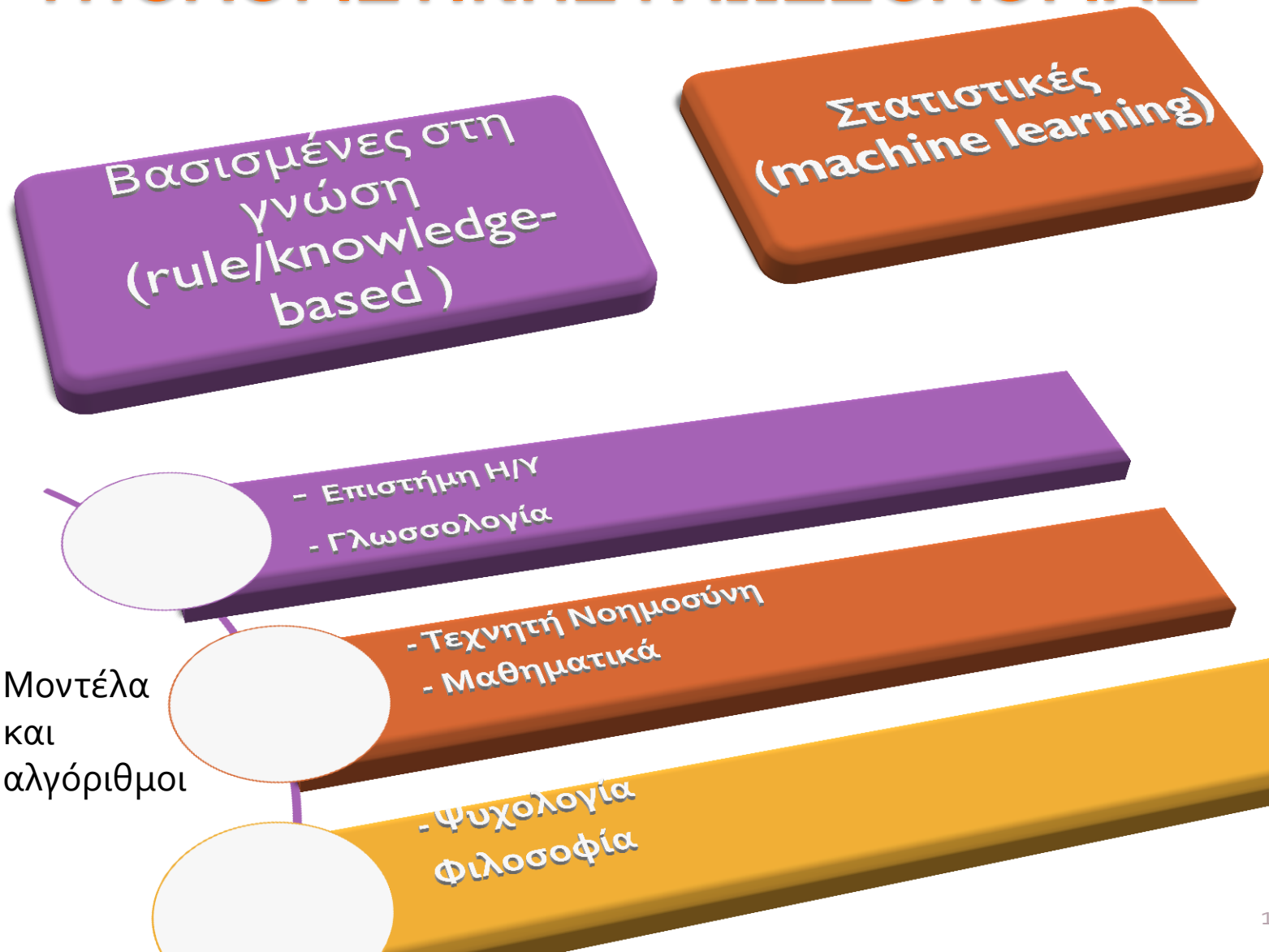
Επεξεργασία Φυσικής Γλώσσας (NLP)

- Διόρθωση ορθογραφίας, εύρεση / εξόρυξη πληροφορίας, μηχανική ή αυτόματη μετάφραση, έλεγχος γραμματικής, συστήματα ερωτοαποκρίσεων.

Αναγνώριση Φωνής (SR)

- Στατιστική επεξεργασία σήματος, κατανόηση φυσικής γλώσσας, νευρωνικά συστήματα, αναγνώριση προτύπων, φωνολογία.

ΜΕΘΟΔΟΙ NLP & ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΓΛΩΣΣΟΛΟΓΙΑΣ



Ερωτήματα Υπολογιστικής Γλωσσολογίας

- Ποια είναι η δομή της γλώσσας και ποιοι μηχανισμοί προσδιορίζουν την παραγωγή της;
- Μπορούμε να αναχθούμε στους μηχανισμούς αυτούς;
- Μπορούμε να δημιουργήσουμε μια μετα-γλώσσα που θα απεικονίζει με ρητό και τυπικό τρόπο αυτούς τους μηχανισμούς;

Έρευνα της Υπολογιστικής Γλωσσολογίας (1)

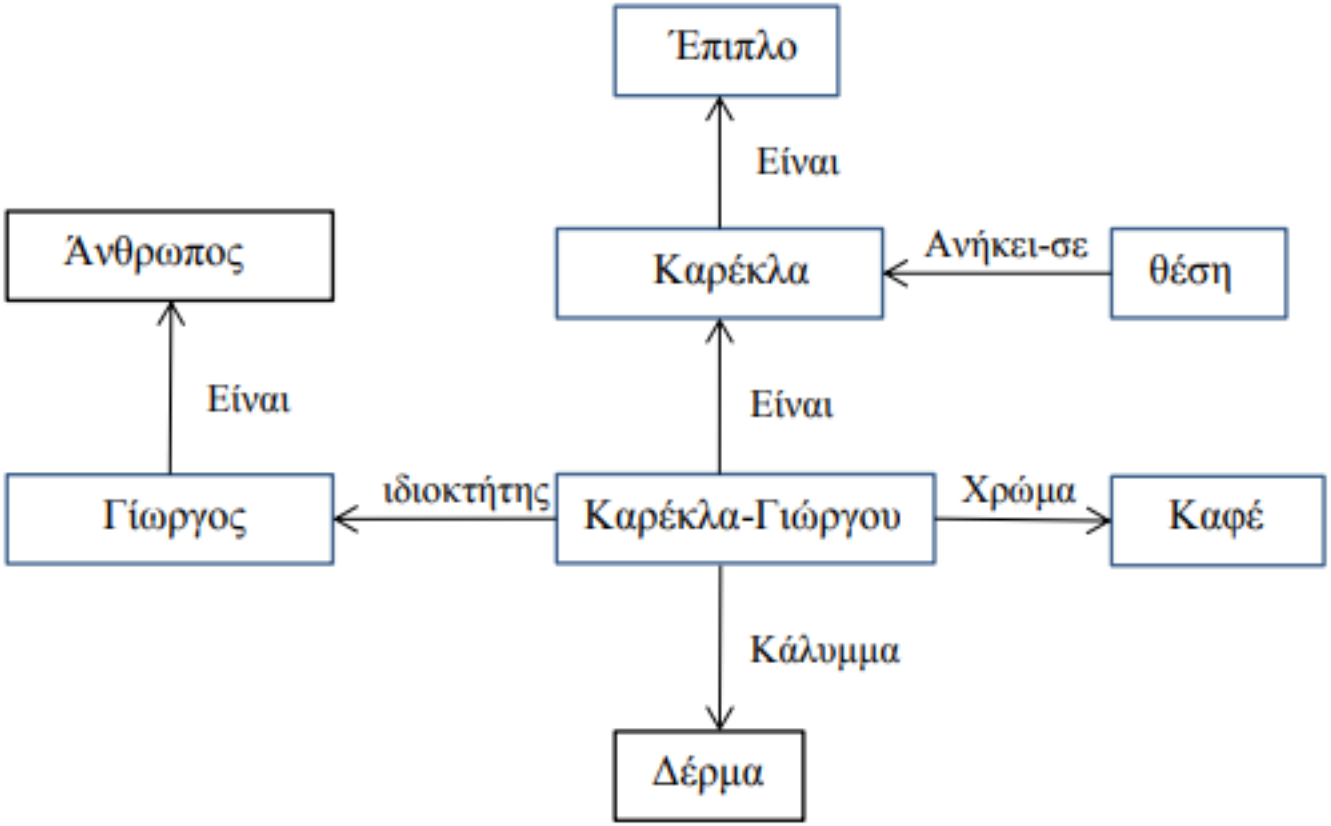
- **Γλωσσικοί Πόροι**.

Στο πεδίο αυτό υλοποιούνται μοντέλα για

1. τις **συνιστώσες της γλώσσας** (φωνήματα, συλλαβές, μορφήματα, λέξεις, φράσεις, προτάσεις, κείμενα) και
2. τα **επίπεδα γλωσσικής ανάλυσης** (φωνολογία, μορφολογία, σύνταξη, σημασιολογία, πραγματολογία).

Η μορφή αυτών των πόρων είναι τα ηλεκτρονικά λεξικά, τα σημασιολογικά δίκτυα, τα δέντρα απόφασης, τα νευρωνικά δίκτυα κ.λπ.

Σημασιολογικά δίκτυα (SEMANTIC NETWORKS)



Σημασιολογικά δίκτυα

- Ένα σημασιολογικό δίκτυο είναι ένα γραφικό σύστημα με σκοπό την αναπαράσταση γνώσης με μορφή συνδεδεμένων κόμβων και τόξων με ετικέτες.
- Πιο συγκεκριμένα χρησιμοποιούνται για προτασιακή αναπαράσταση, για αυτό τα συναντάμε και ως προτασιακά δίκτυα.

Στην προτασιακή λογική ή προτασιακό λογισμό (propositional calculus) ενδιαφερόμαστε για δηλωτικές προτάσεις (declarative sentences) που μπορούν να είναι αληθείς (true) ή ψευδείς (false), αλλά όχι και τα δύο.

Παραδείγματα προτάσεων που εκφράζουν δηλώσεις:

Έχει ήλιο.

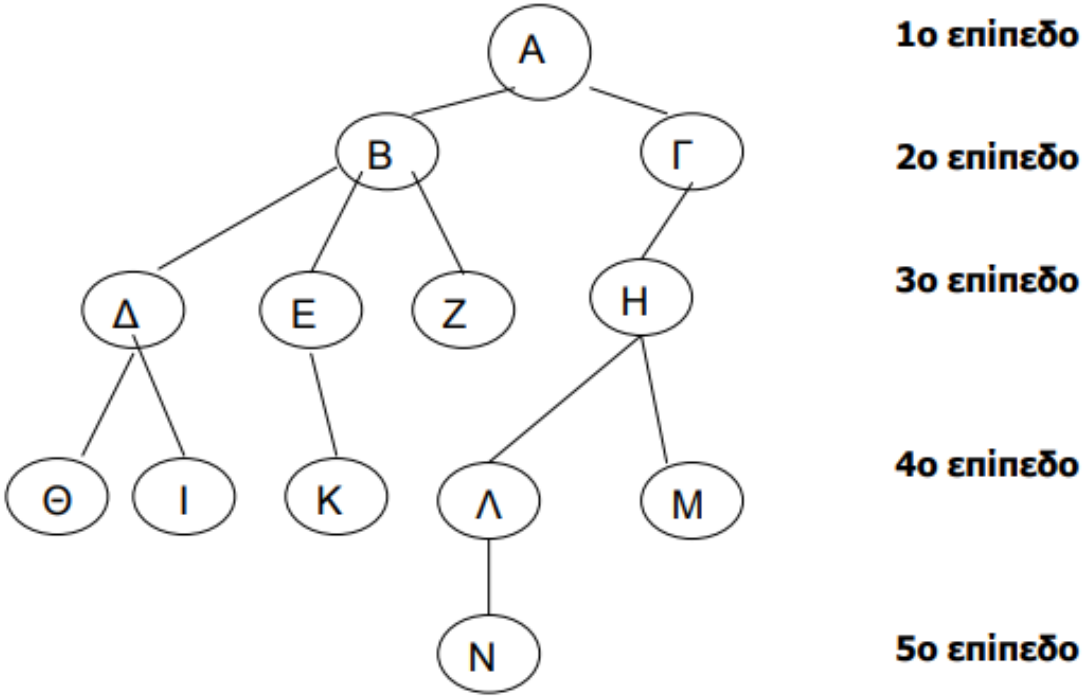
Κάνει ζέστη.

Θα διψάσεις.

- Οι πρώτες υλοποιήσεις σημασιολογικών δικτύων σε υπολογιστές έγιναν για την τεχνητή νοημοσύνη και για μηχανική μετάφραση, όμως νεότερες εκδόσεις τους χρησιμοποιούνταν από πολύ καιρό στη φιλοσοφία, την ψυχολογία και τη γλωσσολογία.

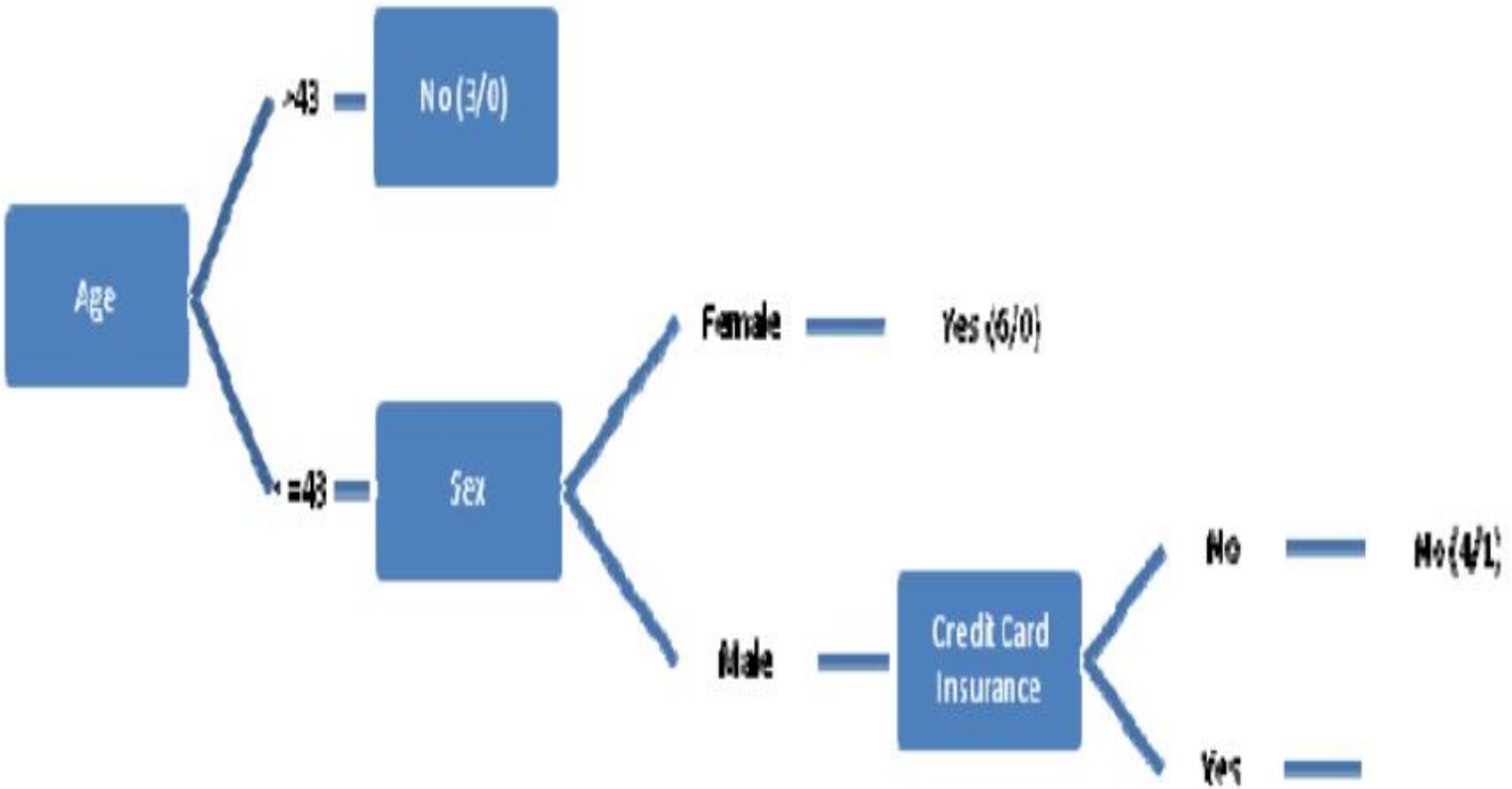
Δέντρα Απόφασης (Decision Trees)

- Τα δέντρα απόφασης είναι μια δημοφιλής δομή για καθοδηγούμενη εκμάθηση.



Παράδειγμα Δέντρου Απόφασης

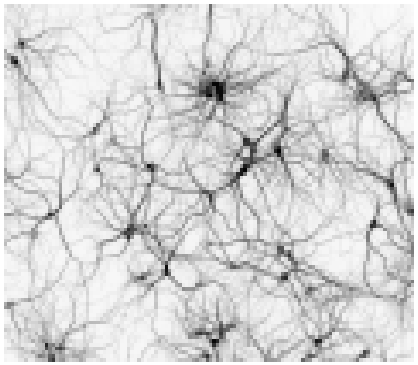
- Ένα δέντρο αποφάσεων τριών κόμβων για τη βάση δεδομένων προώθησης πιστωτικών καρτών



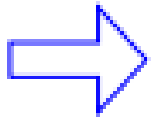
Τι είναι τα Τεχνητά Νευρωνικά Δίκτυα;

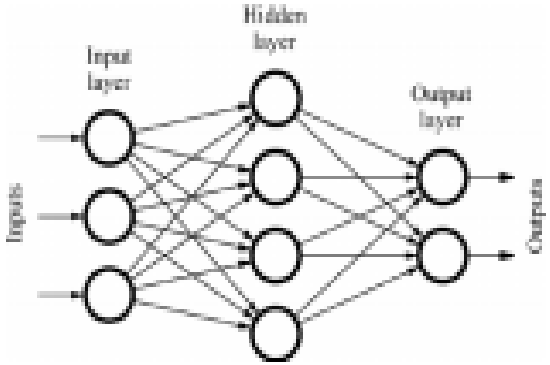
Τεχνητό Νευρωνικό Δίκτυο

Ένα *Τεχνητό Νευρωνικό Δίκτυο* (*Artificial Neural Network*) είναι ένα υπολογιστικό σύστημα υλικού και λογισμικού του οποίου η δομή και η λειτουργία είναι εμπνευσμένη από τον τρόπο λειτουργίας των *Βιολογικών Νευρικών δικτύων*, τα οποία αποτελούν δομικά συστατικά των εγκεφάλων των ζώων και των ανθρώπων.



Βιολογικό Νευρικό Δίκτυο





Τεχνητό Νευρωνικό Δίκτυο

Έρευνα της Υπολογιστικής Γλωσσολογίας (2)

- **Γλωσσικά Εργαλεία Υποδομής**.

Το πεδίο αυτό περιλαμβάνει υπολογιστικά συστήματα για την ανάπτυξη Γλωσσικών Πόρων:

- λεξικογραφικές βάσεις δεδομένων,
- συστήματα διαχείρισης σωμάτων κειμένων,
- συστήματα για τη συγγραφή ή την αυτόματη εκμάθηση
- υπολογιστικά μοντελοποιημένων - γραμματικών κανόνων κ.λπ.

Έρευνα της Υπολογιστικής Γλωσσολογίας (3)

- **Γλωσσικά προϊόντα**. Στα γλωσσικά προϊόντα ανήκουν **τα υπολογιστικά συστήματα που χρησιμοποιούν τους Γλωσσικούς Πόρους** είτε

1. για να ικανοποιήσουν πληροφοριακές ανάγκες των χρηστών (π.χ. εφαρμογές για την περιήγηση σε ηλεκτρονικά λεξικά ή σε γνωσιακές βάσεις δεδομένων, μηχανές αναζήτησης κ.λπ.)
2. είτε για να επεξεργαστούν αυτόματα κείμενο ή ομιλία

Εφαρμογές υπολογιστικής γλωσσολογίας

- τα ηλεκτρονικά λεξικά,
- οι τράπεζες ορολογίας,
- τα σώματα (**corpora**) κειμένων,
- τα συστήματα ελέγχου ορθογραφίας, γραμματικής και ύφους,
- τα συστήματα ανάκτησης πληροφορίας,
- τα συστήματα αναγνώρισης φωνής και μηχανικής μετάφρασης,
- τα μοντέλα διαπροσωπικού και διαλόγου ανθρώπου και υπολογιστή,
- η υπαγόρευση κειμένου στον Η/Υ,
- η έξυπνη οπτική αναγνώριση χαρακτήρων (**OCR**),
- η σύνθεση κειμένου,
- τα συστήματα ελεύθερης αναζήτησης κειμένου με γλωσσική υποστήριξη

ΥΠΑΡΧΟΝΤΑ ΛΟΓΙΣΜΙΚΑ

- Ορθογράφος
- Θησαυρός
- Συλλαβιστής
- Υπολογιστικά Λεξικά

Speller (ορθογράφος)

- Ο ορθογραφικός διορθωτής βασίζεται σε ένα λεξικό 65.000 λημμάτων, από το οποίο παράγονται όλοι οι κλιτοί τύποι (συνολικά περισσότεροι από 1.600.000 τύποι).
- Μια από τις αδυναμίες αυτού του εργαλείου είναι ότι ελέγχει μόνο ορθογραφικά κάθε λέξη χωρίς να εξετάζει τις λέξεις και σε σχέση με το άμεσο περιβάλλον τους, π.χ. το άρθρο με το επίθετο και το ουσιαστικό, το ρήμα με τις αντωνυμίες κ.λπ.

Θησαυρός (Thesaurus)

- ένα ειδικού τύπου λεξικό, το οποίο επιχειρεί να αποδώσει τη σημασία λέξεων ή εκφράσεων της νέας ελληνικής μέσω συνωνύμων, αντωνύμων και παραδειγμάτων χρήσης.
- Οι λημματικοί τύποι, οι σημασίες τους και τα συνώνυμα/αντίθετα (ανά σημασία) συνοδεύονται από υφολογικές και πραγματολογικές πληροφορίες.
- Οι πληροφορίες αυτές όμως αφορούν συγκεκριμένα λήμματα και όχι μορφολογικούς τύπους του ιδίου λήμματος οι οποίοι διαφοροποιούνται, λόγω του ύφους τους, σε διάφορα γλωσσικά περιβάλλοντα.

Συλλαβιστής (hyphenator)

- υποδεικνύει τα σημεία συλλαβισμού μιας λέξης, για να μπορεί το σύστημα στοιχειοθεσίας να συλλαβίσει αυτόματα αυτή τη λέξη, όταν τείνει να υπερβεί το εκτυπώσιμο περιθώριο.
- Οι κανόνες που χρησιμοποιεί χωρίζονται σε δύο κατηγορίες:

α) σε αυτούς που υλοποιήθηκαν με ανθρώπινη επιμέλεια και σε αντιστοιχία με τους κανόνες συλλαβισμού της νέας ελληνικής και

β) σε εκείνους που παρήχθησαν αυτόματα με βάση την πληροφορία συλλαβισμού που υπάρχει στο Μορφολογικό Λεξικό.

- Περιέχει ωστόσο κα περίπου 300 συλλαβισμένους τύπους, αρκετοί από τους οποίους παρουσιάζουν σημασιολογική ασάφεια (π.χ. ή-πι-α, λό-γι-α...)

Υπολογιστικά λεξικά

- α) το υπολογιστικό μορφολογικό και συντακτικό λεξικό της Ν. Ελληνικής που αναπτύχθηκε από το ΙΕΛ. Περιλαμβάνει 20.149 λήμματα κωδικοποιημένα σε μορφολογικό και συντακτικό επίπεδο σύμφωνα με το μοντέλο PAROLE.
- β) Το ανεξάρτητο εργαλείο το Υπολογιστικό Μορφολογικό Λεξικό της νέας ελληνικής γλώσσας (Baltzis, Kolalas & Eumeridou, 2005b).

Τομείς έρευνας Υπολογιστικής Γλωσσολογίας

- η επεξεργασία φυσικής γλώσσας (**N**atural **L**anguage **P**rocessing)
- Η αναγνώριση φωνής (**S**peech **R**ecognition)

NLP

- ασχολείται με τις δομές δεδομένων και τους αλγορίθμους μας φυσικής γλώσσας
- Οι μέθοδοι έρευνας της NLP είναι βασισμένες στη γνώση (**rule/knowledge-based**) ή είναι βασισμένες στη στατιστική (**data-driven**) και χρησιμοποιούν μοντέλα και αλγορίθμους από την επιστήμη των Η/Υ, τη γλωσσολογία, την τεχνητή νοημοσύνη, τη λογική, τα μαθηματικά, την ψυχολογία, τη φιλοσοφία ή τη γνωσιακή επιστήμη

Speech Recognition

- πολύπλοκα συστήματα που εμπεριέχουν ένα μεγάλο σύνολο γνωστικών πεδίων (στατιστική επεξεργασία σήματος, η κατανόηση φυσικής γλώσσας, τα νευρωνικά συστήματα, η αναγνώριση προτύπων, η φωνολογία)
- στοχεύουν στην αναγνώριση ή κατανόηση της φυσικής γλώσσας από υπολογιστή όπως επίσης και στην παραγωγή ή σύνθεση φυσικής γλώσσας

ΙΔΙΑΙΤΕΡΟΤΗΤΕΣ ΓΛΩΣΣΑΣ

Ασάφεια (ambiguity)

- Σημασιολογική (semantic)
- Συντακτική (syntactic)
- Λεξική (lexical)

Ασάφεια & Η/Υ

- κοινοί ορθογραφικοί τύποι > διαφορετικά λήμματα
- Κοινοί ορθογραφικοί τύποι: (διαφορετικά μορφοσυντακτικά χαρακτηριστικά)

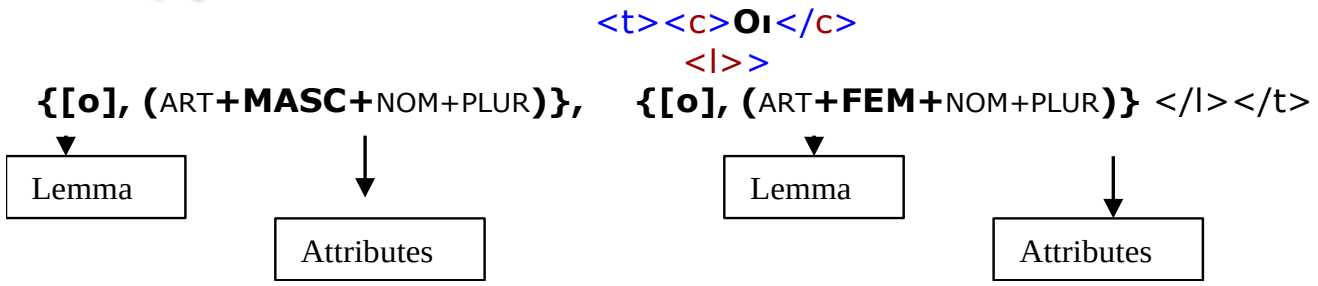
Λεξική ασάφεια

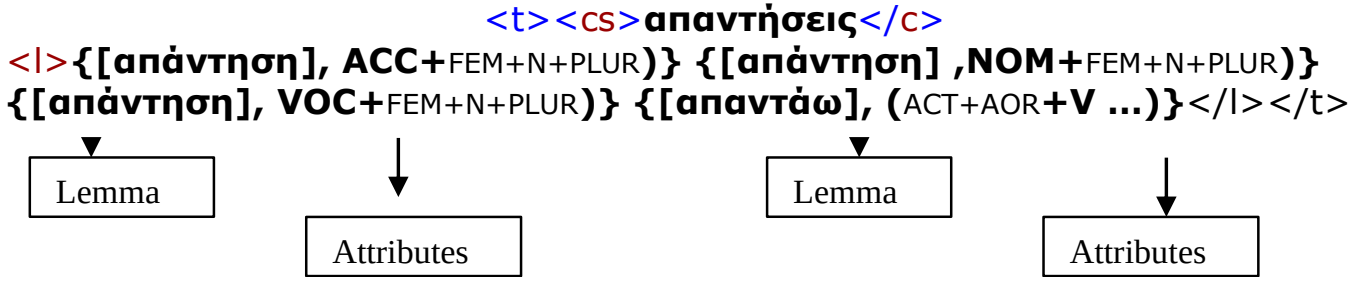
Μέρος του Λόγου	Αριθμός λέξεων
Αριθμός μοναδικών κλιτικών τύπων	873,701
Ασαφείς κλιτικοί τύποι (από διαφορετικά λήμματα)	39,119
Ασαφείς κλιτικοί τύποι (από το ίδιο λήμμα	4,758
Σύνολο ασαφών τύπων	917,578

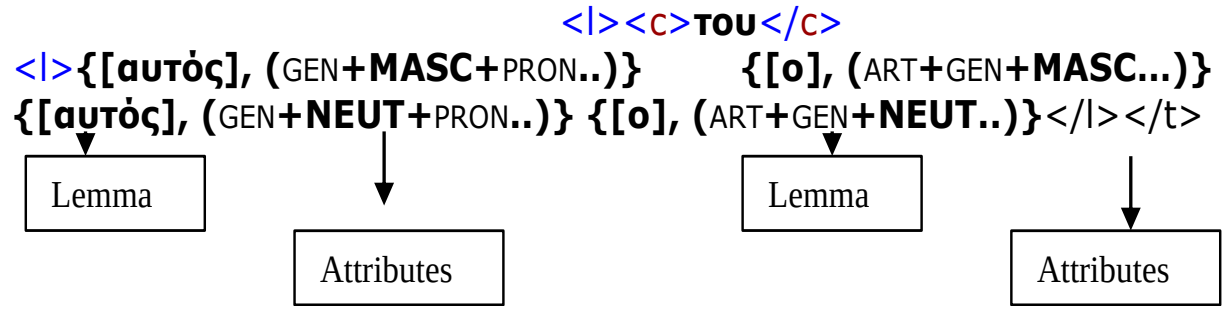
Πίνακας 1. Στατιστικά στοιχεία λεξικής ασάφειας

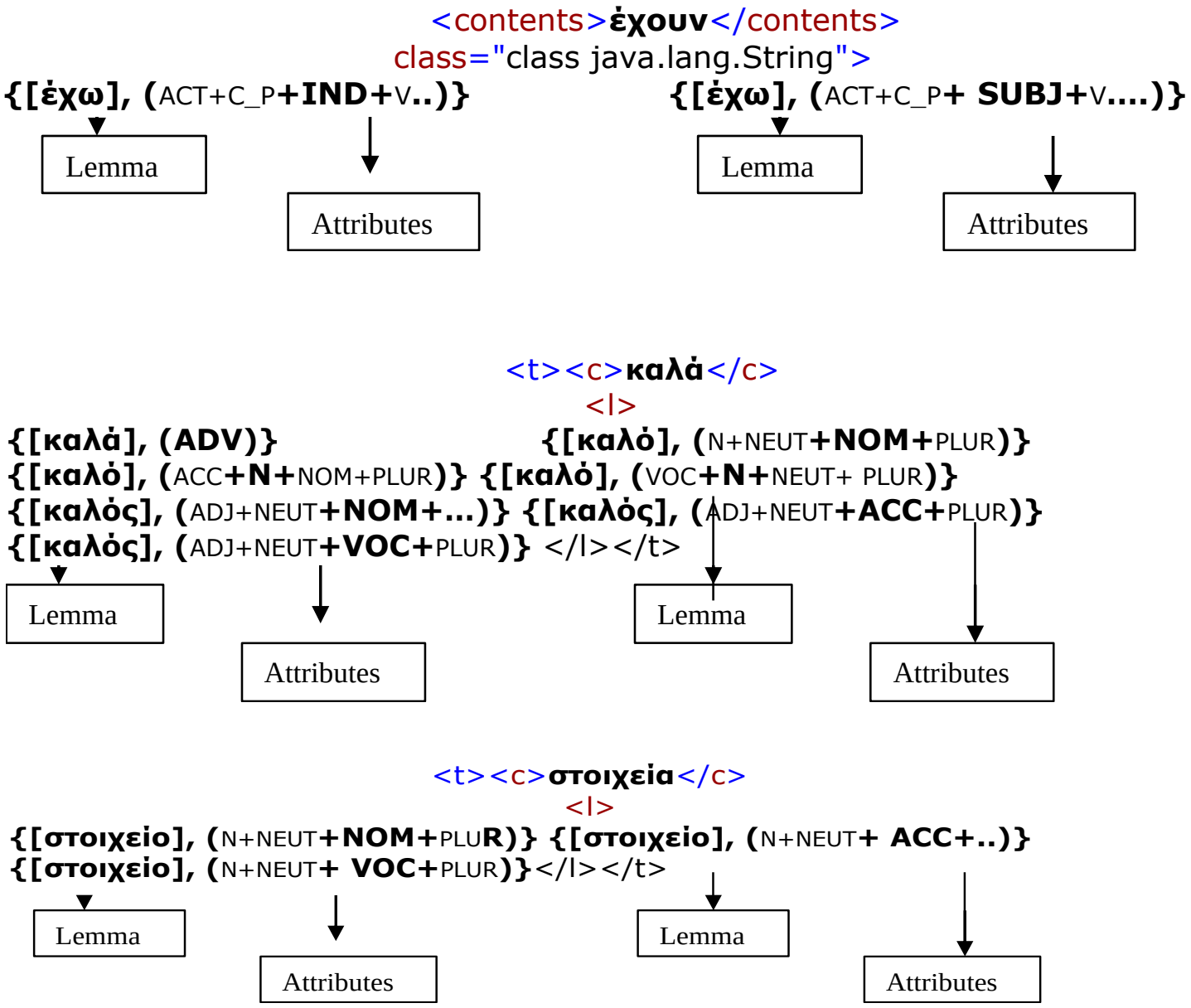
1) Λεξική Ασάφεια

Οι απαντήσεις του έχουν καλά στοιχεία








<contents>έχουν</contents>
class="class java.lang.String">
{[έχω], (ACT+C_P+IND+V..)}
{[έχω], (ACT+C_P+ SUBJ+V....)}
Lemma
Attributes
Lemma
Attributes
<t><c>καλά</c>
<l>
{[καλά], (ADV)}
{[καλό], (N+NEUT+NOM+PLUR)}
{[καλό], (ACC+N+NOM+PLUR)}
{[καλό], (VOC+N+NEUT+ PLUR)}
{[καλός], (ADJ+NEUT+NOM+...)}
{[καλός], (ADJ+NEUT+ACC+PLUR)}
{[καλός], (ADJ+NEUT+VOC+PLUR)} </l></t>
Lemma
Attributes
Lemma
Attributes
<t><c>στοιχεία</c>
<l>
{[στοιχείο], (N+NEUT+NOM+PLUR)}
{[στοιχείο], (N+NEUT+ ACC+..)}
{[στοιχείο], (N+NEUT+ VOC+PLUR)}</l></t>
Lemma
Attributes
Lemma
Attributes

Λεξική ασάφεια

- τα **παροξύτονα ισοσύλλαβα** αρσενικά σε **-ας** και τα **παροξύτονα ισοσύλλαβα θηλυκά -α** που σχηματίζουν κοινούς όλους τους τύπους της κλιτικής τους παραγωγής με διαφορά στο 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 το χαρακτηρισμό (γενική, ενικός, θηλυκό [κεφάλας] < [κεφάλα]), άλλοτε το χαρακτηρισμό (ονομαστική, ενικός, αρσενικό [κεφάλας] < [κεφάλας]).
- Ίδια συμπεριφορά έχουν τα ουσιαστικά [καράφλας, ο], [καράφλα, η], [ηδονοβλεψίας, ο], [ηδονοβλεψία, η] κ.ά. Στην ευρύτερη κατηγορία της λεξικής αμφισημίας ουσιαστικού με ουσιαστικό, τα ονόματα αυτά καταλαμβάνουν το 7,49%.

Lexical ambiguity (examples)

- {**αγωγών** < [αγωγός = conduit], [αγωγή = education]}, {**γραμματικών** < [γραμματικός = secretary], [γραμματική = grammar]}, {**αρωγών** < [αρωγός = helper], [αρωγή = assistance]}, {**αυλών** < [αυλός = pipe], [αυλή = playground]}

Ambiguity in corpora

POS Ambiguity Schemes	Examples words	% occurrence in the corpus
Pronoun-Article	το, τον, τη, την, τις	25,38% (520611 tokens)
Pronoun-Preposition	με, σε	3,78% (77627 tokens)
Adjective-Adverb	λίγο	3,92% (80585 tokens)
Preposition-Particle-Conj	για	2,16% (44295 tokens)
Verb-Noun	ερωτήσεις, γέννα	1,71% (35066 tokens)
Adjective-Adverb-Noun	μέσα, άδεια	1,22% (25040 tokens)
Adjective-Noun	επίπεδο	6,30% (129263 tokens)
Adverb-Conjunction	καθώς	0,63% (12963 tokens)
Pronoun-Adverb	μόνο	0,20% (4305 tokens)
Verb-Adverb	δω	0,08% (1800 tokens)

Ambiguity vs NLP

- Η άρση της ασάφειας αποτέλεσε και αποτελεί πρόκληση στην έρευνα της Γλωσσολογίας και της Υπολογιστικής Γλωσσολογίας.
- 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.

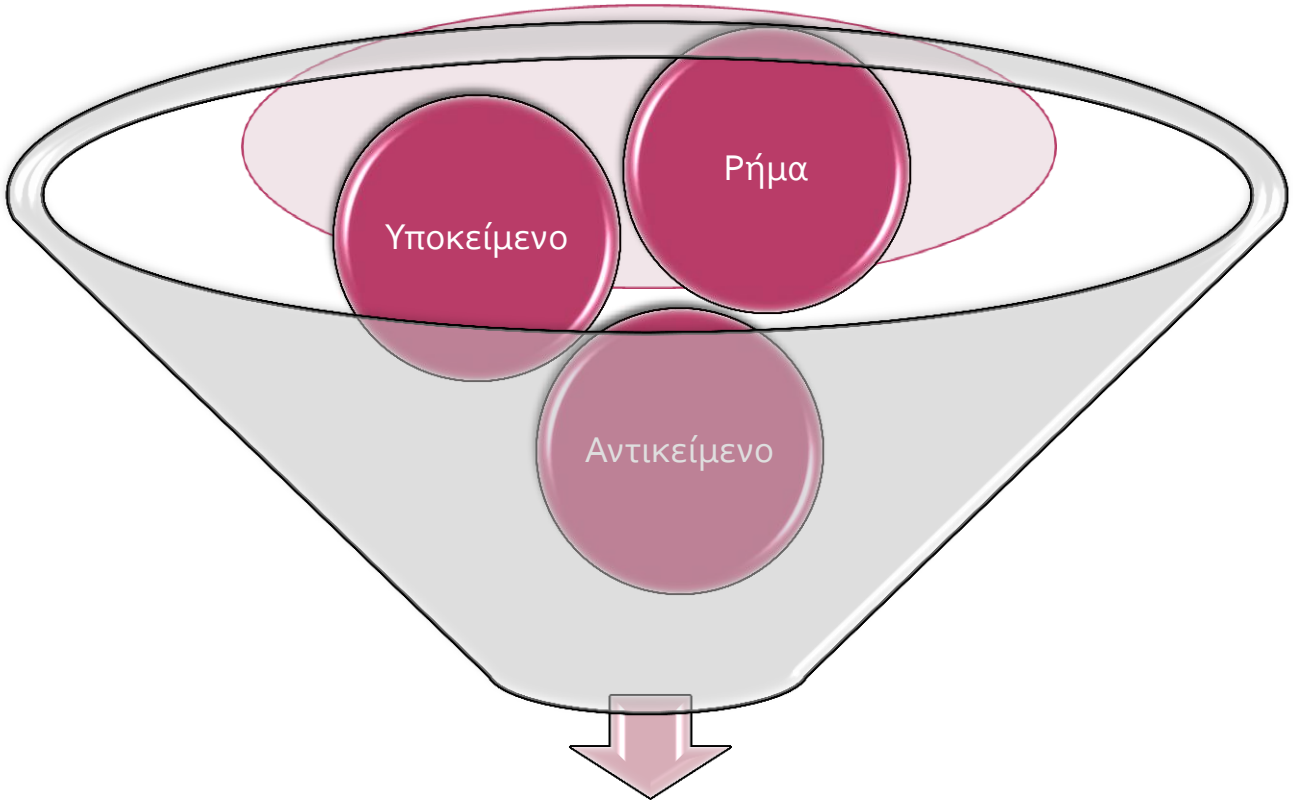
Άρση Αμφισημίας (tagger)

- Οι αμφίσημοι τύποι θα χαρακτηριστούν μορφοσυντακτικά από έναν tagger.
- Επιπλέον ο tagger θα προσπαθήσει να «μαντέψει» τα μορφοσυντακτικά χαρακτηριστικά της λέξης (τουλάχιστον το μέρος του λόγου) ακόμη και για τους λεξικούς τύπους που δεν χαρακτηρίζονται από κανένα μορφοσυντακτικό χαρακτηρισμό (attribute).
- Αυτό θα γίνει αν εξετάσει κυρίως το γλωσσικό περιβάλλον της (τις λέξεις που **προηγούνται ή/και έπονται**).
- Κατ' αυτό τον τρόπο θα ολοκληρωθεί η λειτουργικότητα του λεξικού και η μετέπειτα ανάλυση και εξαγωγή της μορφοσυντακτικής πληροφορίας θα στηρίζεται σε αληθή δεδομένα.

Παράδειγμα ασάφειας

- η λέξη [*απαντήσεις* < **απάντηση**] μπορεί: α) να παίξει τον ρόλο της κεφαλής σε μια ονοματική φράση, β) να παίξει τον 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.

2) «Ελευθερία» Μετακίνησης όρων (free word order)



Ο Γιώργος αγαπά τη Μαρία

Ελευθερία μετακίνησης όρων (free-word-order)

- οι χαλαροί συνεκτικοί δεσμοί των συστατικών στοιχείων της πρότασης. Αυτό σημαίνει ότι τα συστατικά στοιχεία της πρότασης (ονοματική φράση ή ρηματική φράση ή προθετική φράση κ.λπ.) όπως επίσης και τα δομικά στοιχεία των συστατικών αυτών (ουσιαστικό ή άρθρο ή ρήμα κ.λπ.) δεν υπόκεινται σε κανόνες, αλλά είναι δυνατό να καταλάβουν διάφορες θέσεις μέσα στην πρόταση θέση.
- Η φράση [*Ο Γιώργος αγαπά τη Μαρία*] μπορεί να δηλωθεί με 6 τρόπους:

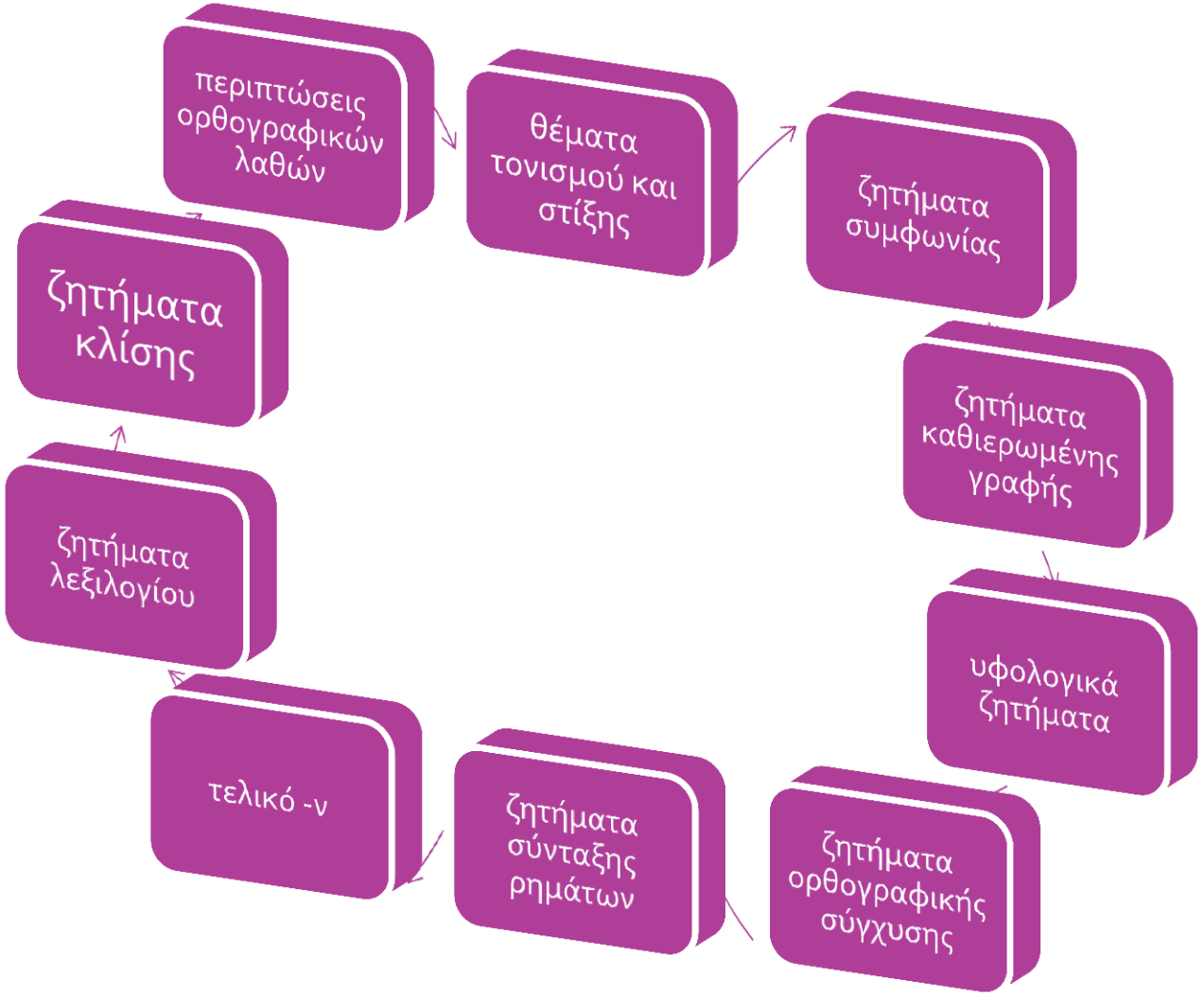
- [*αγαπά τη Μαρία ο Γιώργος*]
- [*ο Γιώργος τη Μαρία αγαπά*]
- [*τη Μαρία ο Γιώργος αγαπά*]
- [*τη Μαρία αγαπά ο Γιώργος*]
- [*αγαπά ο Γιώργος τη Μαρία*].

- Το στοιχείο αυτό κάνει πιο πολύπλοκη την υπολογιστική επεξεργασία της γλώσσας

Γραμματικός Διορθωτής (Grammar checker)

- προσφέρει τη δυνατότητα επιλογής επιπέδου γλώσσας ή κειμενικού είδους με σκοπό να καλύπτει ακόμη και τις πιο εξειδικευμένες ανάγκες των χρηστών
- προσφέρει τη δυνατότητα επιλογής επιπέδου γλώσσας ή κειμενικού είδους με σκοπό να καλύπτει ακόμη και τις πιο εξειδικευμένες ανάγκες των χρηστών.

GRAMMAR CHECKER



Υπολογιστική Γλωσσολογία (σήμερα)

- ο άκρως διεπιστημονικός τομέας της ***Επεξεργασίας της Φυσικής Γλώσσας (Natural Language Processing-NLP)*** αναπτύχθηκε σε σημείο που σήμερα είναι μέρος της παγκόσμιας οικονομίας (Google)
- Ο όρος NLP δεν είναι ο μόνος με τον οποίο ο τομέας είναι γνωστός: εξίσου χρήσιμοι είναι και οι όροι ***Γλωσσική Τεχνολογία (Human Language Technology-HLT)*** και ***Υπολογιστική Γλωσσολογία (Computational Linguistics-CL)***.

Δημοφιλείς και λιγότερο γνωστές εφαρμογές της Υπολογιστικής Γλωσσολογίας

- Google (η αναζήτηση γίνεται με λέξεις-κλειδιά (keywords) τις οποίες δίνετε εσείς στη μηχανή αναζήτησης του Google και η μηχανή κάνει αυτό που αποκαλούμε στην Υπολογιστική Γλωσσολογία *ανάκτηση πληροφορίας (information extraction)*σε μια σχετικά απλή μορφή

Δημοφιλείς και λιγότερο γνωστές εφαρμογές της Υπολογιστικής Γλωσσολογίας

- κινητό (έχει ενσωματωμένη κάποια ελαφριά τεχνολογία σύνθεσης φωνής)

- ομιλούσες ιστοσελίδες, ομιλούντα αυτοκίνητα και συστήματα διαχείρισης των οικιακών συσκευών με τα οποία διαλέγεται ο χρήστης προφορικά και όχι γραπτά

Εφαρμογές Υπολογιστικής Γλωσσολογίας

- Πίσω στο κείμενο: γράφετε στο Word και αυτό σας κάνει ορθογραφικό έλεγχο (spelling checking) με το να κοκκινίζει λέξεις και να σας προτείνει και λύσεις. Προφανώς, μπορεί και αναγνωρίζει λέξεις, τις συγκρίνει με κάποια πρότυπα που έχει αποθηκευμένα και, με βάση την ομοιότητα και ίσως και συγκειμενικές πληροφορίες, προχωρά σε συγκεκριμένες ενέργειες.

Εφαρμογές Υπολογιστικής Γλωσσολογίας

- Οι υπολογιστές ήδη καλούνται να κάνουν αυτόματες περιλήψεις κειμένων και να παράγουν νέο κείμενο από δεδομένα που δεν είναι αναγκαστικά κειμενικά.
- Στην περίπτωση των δελτίων καιρού το κείμενο παράγεται από καθαρά αριθμητικά δεδομένα και σε διαφορετικές γλώσσες ταυτόχρονα, με κλασικό παράδειγμα το δελτίο καιρού του Καναδά

Εφαρμογές Υπολογιστικής Γλωσσολογίας

- Εταιρείες δημοσκοπήσεων καλούνται, έναντι αμοιβής φυσικά, να εκτιμήσουν τον αντίκτυπο που είχαν στην κοινωνία διαφημίσεις, πολιτικά γεγονότα κ.τλ. Αυτό το κάνουν αναλύοντας τεράστιες ποσότητες κειμενικών, κυρίως, δεδομένων για να εντοπίσουν και να αξιολογήσουν τις γνώμες που διατυπώνονται σχετικά με το θέμα που ενδιαφέρει κάθε φορά.

Μερικά ωραία προβλήματα της Υπολογιστικής Γλωσσολογίας !!

- *Μηχανική Μετάφραση* (Machine Translation): Ορισμένοι πιστεύουν ότι η Μηχανική Μετάφραση είναι το βασικό πρόβλημα της Υπολογιστικής Γλωσσολογίας.
- Η μετάφραση από γλώσσα σε γλώσσα έχει τεράστιο οικονομικό ενδιαφέρον αλλά προσκρούει στις μεγάλες διαφορές μεταξύ των γλωσσών και στην αμφισημία που χαρακτηρίζει την γλώσσα γενικά.

Μερικά ωραία προβλήματα της Υπολογιστικής Γλωσσολογίας !!

- *Επίλυση Σημασιολογικής Αμφισημίας* (Word Sense Disambiguation): Τι σημαίνει η λέξη 'έφαγε' στο κείμενο «*τον έφαγε η θάλασσα*»; Αν το 'τον' αναφέρεται σε βράχο, το 'έφαγε' μάλλον αναφέρεται στη διάβρωση, αν το 'τον' αναφέρεται σε ναυτικό, το 'έφαγε' μπορεί να σημαίνει πνιγμό ή ακραία ταλαιπωρία με συνέπειες.
- **Και αν θέλουμε να εφαρμόσουμε κάποια μηχανή μηχανικής μετάφρασης σε αυτό το κείμενο πώς ξέρουμε τι σημασία έχει το 'έφαγε';**

Μερικά ωραία προβλήματα της Υπολογιστικής Γλωσσολογίας !!

- *Επίλυση Συντακτικών Αμφισημιών*: Πώς στο ακόλουθο κείμενο ξέρει η μηχανή ότι το 'τον' αναφέρεται στον βράχο και όχι στον ναυτικό; «*Ο ναυτικός ακουμπούσε στον βράχο που ήταν ετοιμόρροπος, γιατί τον είχε φάει τελείως η θάλασσα*».

Μερικά ωραία προβλήματα της Υπολογιστικής Γλωσσολογίας !!

- *Επίλυση αναφορών* (Anaphora resolution): Σε τι αναφέρεται το αντωνυμικό 'αυτό' στο κείμενο «*Η οικονομία μας σημείωσε βελτίωση το τελευταίο εξάμηνο. Αυτό βελτίωσε το κλίμα στις σχέσεις μας με την Ευρώπη*»;

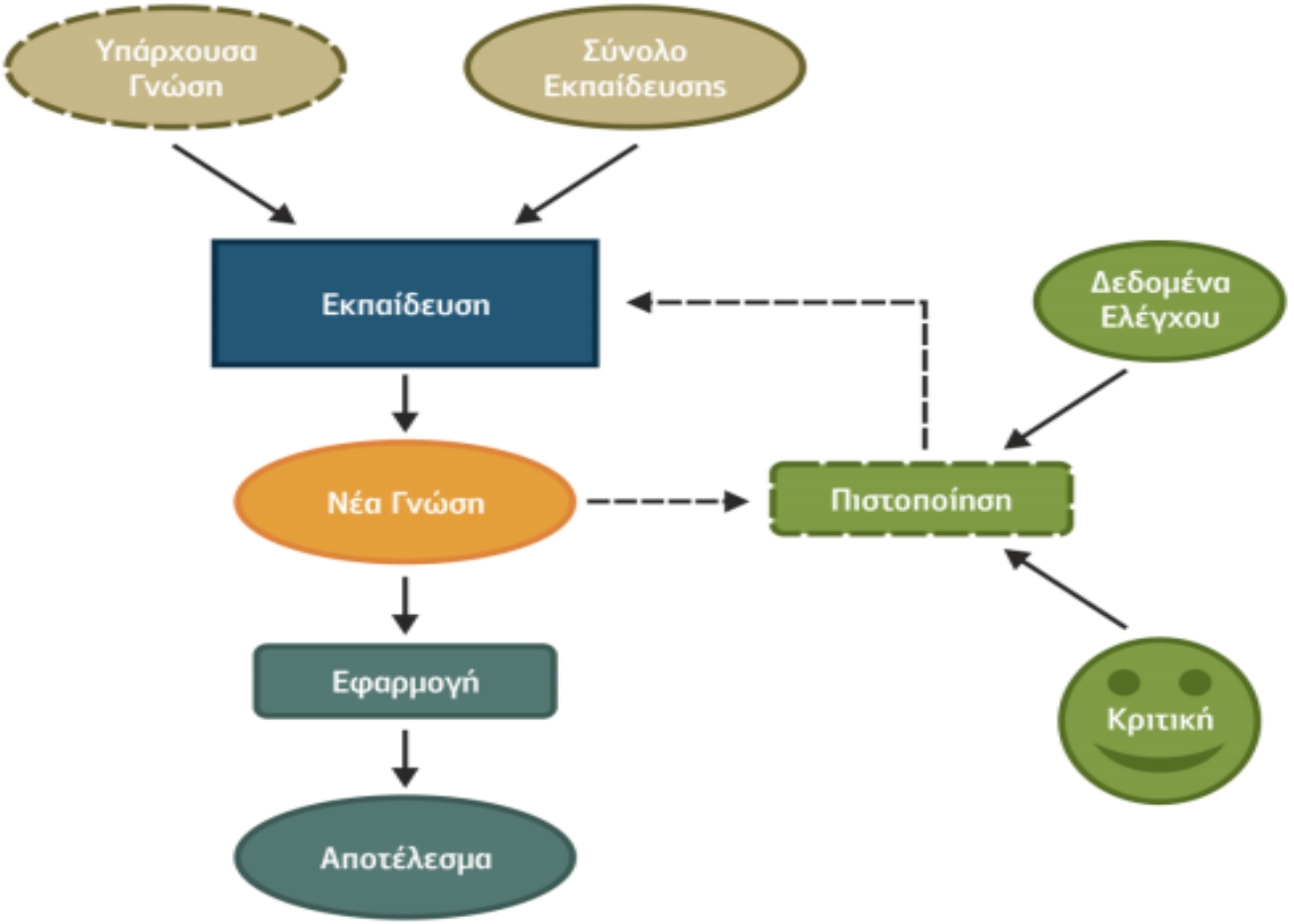
Machine Learning (μηχανική μάθηση)

- Πώς, λοιπόν, θα μπορούσαν οι επιστήμονες του χώρου της ΥΓ να δημιουργήσουν υπολογιστικά συστήματα ικανά να μάθουν, να επιτύχουν, δηλαδή, τη λεγόμενη Μηχανική Μάθηση (Machine Learning);
- Αυτή μπορεί να οριστεί ως το φαινόμενο κατά το οποίο ένα σύστημα βελτιώνει την απόδοσή του κατά την εκτέλεση μιας συγκεκριμένης εργασίας, χωρίς να υπάρχει ανάγκη να προγραμματιστεί εκ νέου.

Μηχανική μάθηση

- η Μηχανική Μάθηση έχει ως σκοπό τη δημιουργία μηχανών ικανών να μαθαίνουν, να βελτιώνουν, δηλαδή, την απόδοσή τους σε κάποιους τομείς μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας.

MACHINE LEARNING



Μηχανική Μάθηση

- Υπάρχουν τρεις βασικές κατηγορίες μηχανικής μάθησης:

- **Μάθηση με επίβλεψη** (supervised learning)
  - μάθηση από παρατήρηση του input και output παραδειγμάτων
- **Μάθηση χωρίς επίβλεψη** (unsupervised learning) ή μάθηση από παρατήρηση
  - μάθηση χωρίς να ξέρουμε το output παραδειγμάτων
- **Ενισχυτική μάθηση** (reinforcement learning)
  - μάθηση μέσω ενίσχυσης (επιβράβευσης)

Τεχνητή νοημοσύνη

- Στο ερώτημα *«Τι είναι Τεχνητή Νοημοσύνη;»* οι ερευνητές του χώρου δίνουν πολλές διαφορετικές απαντήσεις, φαινόμενο που δεν απαντά σε άλλους επιστημονικούς χώρους, όπως η Φυσική, η Χημεία, η Ιατρική κ.ά. Ωστόσο, όλοι φαίνεται να συμφωνούν πως η ΤΝ **είναι επιστήμη και όχι απλώς ένας κλάδος της τεχνολογίας λογισμικού**.
- Κατά τον Patrick Winston (1992), διευθυντής του εργαστηρίου ΤΝ του Πανεπιστημίου MIT, πρωταρχικός σκοπός της ΤΝ είναι *«να κάνει τις μηχανές πιο έξυπνες»*· σε αυτό συμφωνούν οι περισσότεροι από τους ερευνητές που αντιμετωπίζουν την ΤΝ ως αναζήτηση μεθόδων οι οποίες θα κάνουν τους ηλεκτρονικούς υπολογιστές πιο έξυπνους και, συνεπώς, πιο χρήσιμους από όσο είναι σήμερα

Τεχνητή νοημοσύνη (ορισμός)

«Τεχνητή Νοημοσύνη είναι εκείνος ο κλάδος της επιστήμης των υπολογιστών που ασχολείται με το σχεδιασμό ευφυών υπολογιστικών συστημάτων, δηλαδή συστημάτων με χαρακτηριστικά τα οποία σχετίζονται με την ευφυΐα στην ανθρώπινη συμπεριφορά (μάθηση, αιτίαση, επίλυση προβλημάτων, κατανόηση φυσικής γλώσσας, αναγνώριση αντικειμένων κτλ.).»

Προσεγγίσεις στην Τεχνητή Νοημοσύνη

Συστήματα που σκέφτονται όπως ο άνθρωπος «Η αυτοματοποίηση λειτουργιών που σχετίζονται με την ανθρώπινη σκέψη, όπως η λήψη αποφάσεων, η επίλυση προβλημάτων, η μάθηση...»	**Συστήματα που σκέφτονται ορθολογικά** «Η μελέτη νοητικών ικανοτήτων με την χρήση υπολογιστικών μοντέλων»
Συστήματα που δρουν όπως ο άνθρωπος «Η τέχνη της δημιουργίας μηχανών που κάνουν λειτουργίες, οι οποίες, όταν πραγματοποιούνται από ανθρώπους, απαιτούν νοημοσύνη»	**Συστήματα που δρουν ορθολογικά** «Η μελέτη της σχεδίασης ευφυών πρακτόρων»

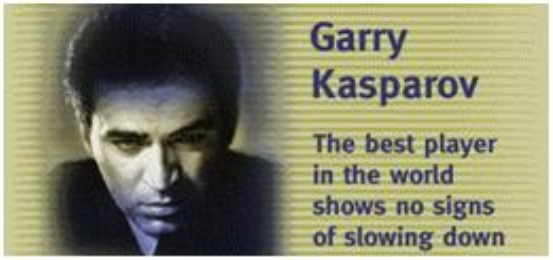
Η ΤΝ σήμερα (1986 -)

- Η Τ.Ν. μετεξελίχθηκε σε επιστήμη:
  - Τα επιτεύγματα στηρίζονται πλέον σε αυστηρές θεωρίες.
  - Νέα "επιτεύγματα" γίνονται αποδεκτά μόνο κατόπιν αυστηρής απόδειξης ή εξαντλητικής πειραματικής επιβεβαίωσης.
  - Το διαδίκτυο αποτελεί ιδανικό χώρο πραγματικής δοκιμής των νέων τεχνολογιών.

Εφαρμογές ΤΝ

Παιχνίδια (Game Playing)

- Ο υπολογιστής της IBM Deep Blue έγινε ο πρώτος που νίκησε τον παγκόσμιο πρωταθλητή στο σκάκι Garry Kasparov με σκορ 3.5-2.5, σε αγώνα επίδειξης το 1997.
- http://www.research.ibm.com/deepblue/



Εφαρμογές ΤΝ

Αυτόνομος έλεγχος (Autonomous control)

- Το ALVINN είναι ένα σύστημα μηχανικής όρασης το οποίο μαθαίνει να οδηγεί ένα όχημα, παρακολουθώντας έναν άνθρωπο-οδηγό.
- Το ALVINN αποτελείται από ένα νευρωνικό δίκτυο ενός κρυμμένου στρώματος.
- Το σύστημα δέχεται είσοδο από τις κάμερες πάνω στο όχημα.
- Η έξοδος είναι η κατεύθυνση προς την οποία πρέπει να κινηθεί το όχημα για να παραμείνει μέσα στο δρόμο.
- Ταξίδευσε 2850 μίλια στις ΗΠΑ διατηρώντας αυτονομία στο 98% του ταξιδιού!

http://www-2.cs.cmu.edu/afs/cs/project/alv/www/navlab_home_page.html

Εφαρμογές ΤΝ

Κατανόηση Λόγου

- Verbmobil: Σύστημα αυτόματης μετάφρασης, εξαρτημένης από τα συμφαζόμενα, σε πραγματικό χρόνο.
  - Γλώσσες: Γερμανικά-Αγγλικά-Ιαπωνικά
  - http://verbmobil.dfki.de/

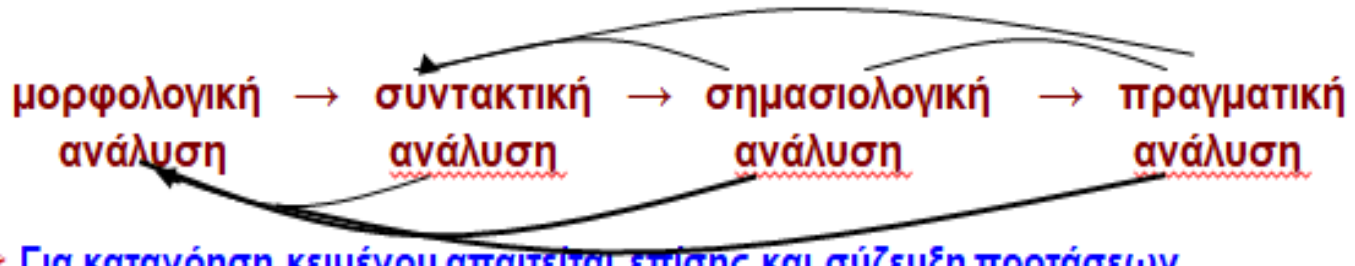
Κατανόηση φυσικής γλώσσας
Το πρόγραμμα ELIZA
(Turing Test του J. Weizenbaum, 1960) 1/2

- **Το πρόγραμμα ELIZA προσομοιώνει ένα Ψυχίατρο που συνομιλεί με μια ασθενή:**
  - MEN ARE ALL ALIKE : ασθενής
  - IN WHAT WAY : μηχανή - ψυχίατρος
  - THEY'RE ALWAYS BUGGING US ABOUT SOMETHING OR OTHER
  - CAN YOU THINK OF A SPECIFIC EXAMPLE?
  - WELL, MY BOYFRIEND MADE ME COME HERE
  - YOUR BOYFRIEND MADE YOU COME HERE
  - HE SAYS I'M DEPRESSED MOST OF THE TIME
  - I'M SORRY TO HEAR YOU ARE DEPRESSED
  - IT'S TRUE, I AM UNHAPPY
  - DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY?
  - I NEED SOME HELP, THAT MUCH SEEMS CERTAIN
  - WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP?
  - PERHAPS I COULD LEARN TO GET ALONG WITH MY MOTHER
  - TELL ME MORE ABOUT YOUR FAMILY

Προβλήματα στην κατανόηση της Φυσικής Γλώσσας

- ✓Cinderella attended the ball
  - ✓Cinderella kicked the ball
- ✓The house was built by the workmen
  - ✓The house was built by the river
- ✓The soldiers shot the prisoners and I saw several fall
  - ✓The soldiers shot the prisoners and I saw several reload
- ✓Η Ελένη είδε την ταινία με τον Δημήτρη
  - ✓Η Ελένη είδε την ταινία με τον Σαρλώ
- ✓Ο Γιώργος είδε τον Νίκο στην πλατεία με τα κιάλια
  - ✓Ο Γιώργος είδε τον Νίκο στην πλατεία με το άγαλμα
- ✓Ο Γιώργος χτύπησε το σκύλο του
  - ✓Ο Γιώργος χτύπησε το πόδι του
- ✓Όλα τα βιβλία και τα περιοδικά που λείπουν κοστίζουν 200.000 δρχ.
  - ✓Όλα τα βιβλία και τα περιοδικά που εκδίδονται ανά τρίμηνο κοστίζουν 200.000 δρχ.

- Η κατανόηση της φυσικής γλώσσας απαιτεί πολύ γνώση.
- Απαιτούνται τέσσερα στάδια με αναδράσεις μεταξύ τους (για προτάσεις):



- Για κατανόηση κειμένου απαιτείται επίσης και σύζευξη προτάσεων

Ευχαριστώ για την προσοχή σας!