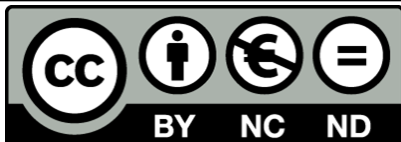


# ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

## ΕΙΣΑΓΩΓΗ



Ευρωπαϊκή Ένωση  
Ευρωπαϊκό Κοινωνικό Ταμείο



Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



# Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.

Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Ιόνιο Πανεπιστήμιο**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.



# Άδειες Χρήσης

---

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons



Ο προπτυχιακός φοιτητής Σπύρος Σαγιαδινός στην παρουσίαση της εργασίας που εκπόνησε η ομάδα του στο μάθημα, στο Πανελλήνιο Συνέδριο Τεχνητής Νοημοσύνης, SETN, Μάιος 2014



# Τι είναι Γλωσσική Τεχνολογία;

---

- Η ανάπτυξη υπολογιστικών μοντέλων επεξεργασίας πληροφορίας (γνώσης) που βρίσκεται σε φυσική (ανθρώπινη) γλώσσα (Σ. Σταματάτος-Παν/μιο Αιγαίου) .
  - Προφορικός λόγος – Επεξεργασία ομιλίας
  - Γραπτός λόγος – Επεξεργασία φυσικής γλώσσας
  - Εικόνα εγγράφου - Οπτική Αναγνώριση Χαρακτήρων

# Γιατί είναι σημαντική;

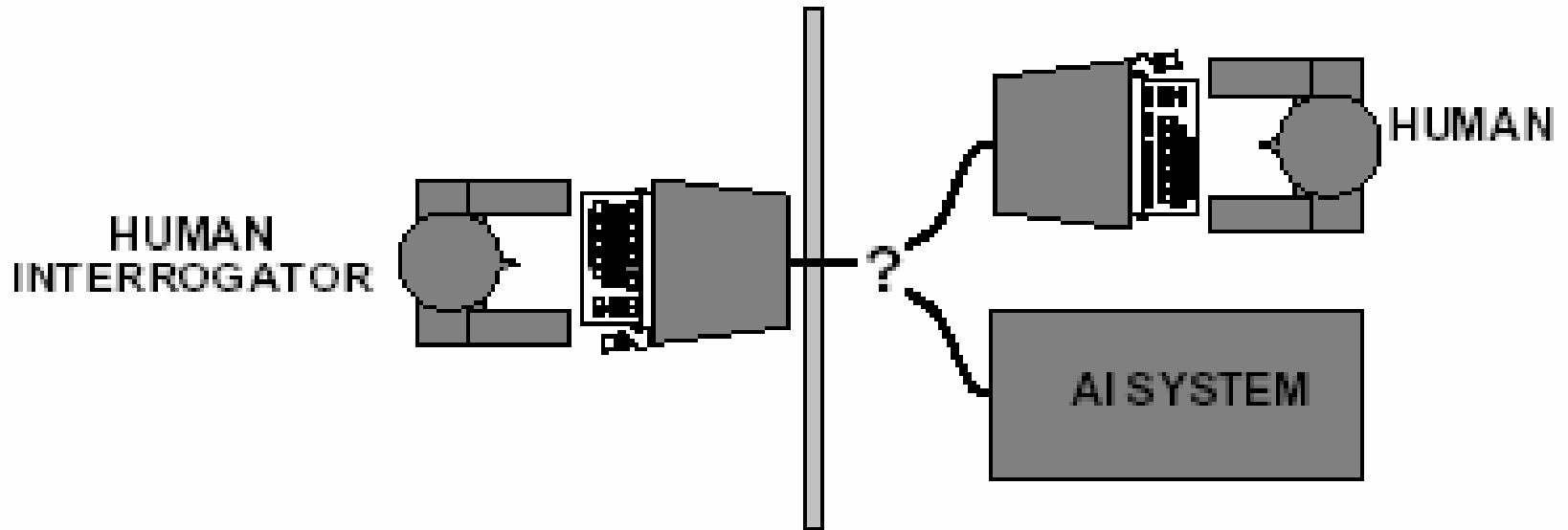
---

- Για την ανάπτυξη πιο εύχρηστων και χρήσιμων υπολογιστικών συστημάτων
  - Διεπαφή σε φυσική γλώσσα
  - Συστήματα κατανόησης ομιλίας
  - Καλύτερα συστήματα ανάκτησης/εξαγωγής πληροφορίας από κείμενα (WWW)
- Για την ανάπτυξη συστημάτων μηχανικής μετάφρασης
- Για την καλύτερη κατανόηση της ανθρώπινης επικοινωνίας/έκφρασης

# Τεστ Turing

---

---



# Computer chatbot 'Eugene Goostman' passes the Turing test

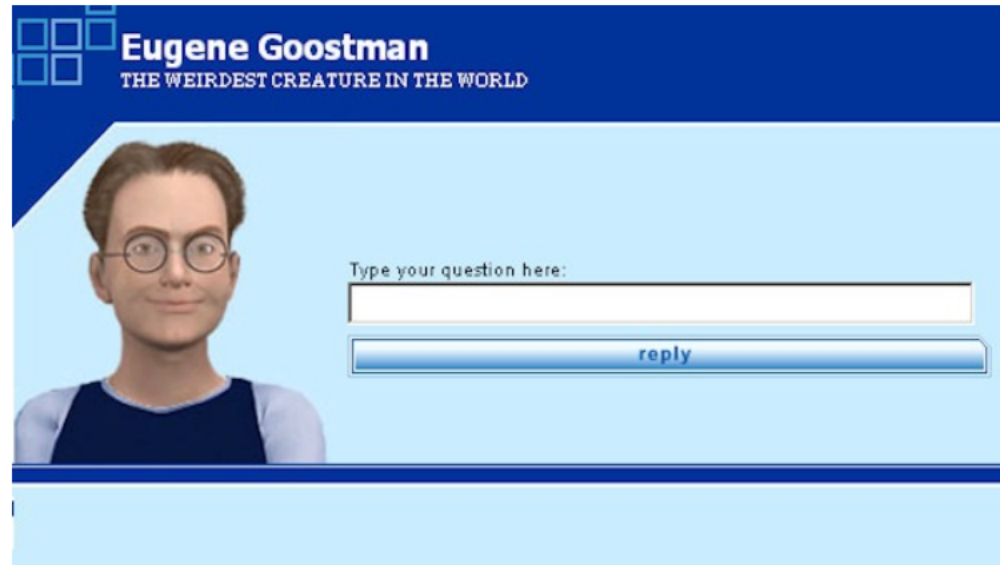
**Summary:** A computer program that pretends to be a 13-year-old Ukrainian boy passed a Turing test at the Royal Society in London yesterday on the 60th anniversary of Turing's death.



By Jack Schofield for Jack's Blog | June 8, 2014 -- 17:00 GMT (10:00 PDT)

[Follow @jackschofield](#)

[Get the ZDNet Insights newsletter now](#)



A computer program that pretends to be a 13-year-old Ukrainian boy called Eugene Goostman passed a Turing test at the Royal Society in London yesterday (Saturday 6 June) by convincing 33 percent of the judges that it was human during a five-minute typed conversation. The test was suggested by computer scientist Alan Turing in 1950, and the competition was held on the 60th anniversary of his death.

27 Comments

5 Votes

1736

262

Tweet

120

Share

more +

Samsung My KN

Take control of your mobile device for work and play

GET IT ON Google

Mastering Google Voice primer for users and business

Everything you need to know to get the most out of Google Voice, in a single bundled package.

[Start reading now. »](#)

- Related Ads**
- 1 Mobile Apps
  - 2 Unlock Your Phone
  - 3 Digital Computer
  - 4 Loebner Prize Contest
  - 5 Alan Turing
- 8



## Test 1

[12:43:37] Judge: Hi, I'm Cathy.

[12:43:43] Eugene: Cathy! Nice to meet you! You're a woman, if I'm not mistaken.

[12:44:20] Judge: Right. What is your name?

[12:44:29] Eugene: People call me Eugene. Or "Zhenya". It is a short name for Eugene in Russia and Ukraine.

[12:45:17] Judge: Nice name. Are you from either area?

[12:45:41] Eugene: Agree :- ) Oh, what a fruitful conversation;-)

[12:45:52] Judge: Do you really think so?

[12:46:04] Eugene: No, your guess was wrong. I don't think so. By the way, I still don't know your specialty – or, possibly, I've missed it?

[12:46:56] Judge: What speciality? What's yours?

[12:47:01] Eugene: Huh? And I forgot to ask you where you are from.

[12:47:51] Judge: The universe and you?

[12:48:01] Eugene: Is it your hobby to ask little poor Jewish boys such tricky questions they don't have an answer for?

[16:12:00] Eugene: I wonder what Eminem's nick mean. I associate it only with a word "enema". But he is a cool guy, I like his The Real Slim Shady.

[16:12:55] Judge: i know what his nickname means, his real name is Marshall Mathers. From the initials M&M phynetically sounding like Eminem.

[16:13:06] Eugene: Eminem is a famous crapper. No, rapper... ripper... Yeah, I mean – rapper. I like his Stan and The Real Slim Shady.

[16:13:32] Judge: I like Hailies Song.

[16:13:46] Eugene: Interesting tastes. I like science-fiction books, by the way. Where do you came from, by the way? Could you tell me about the place where you live?

[16:14:22] Judge: Epsom, home of the derby. yourself?

[16:14:27] Eugene: My guinea pig says that name Derby sounds very nice.

[16:14:41] Judge: i shopuld [sic] talk with him sometime.

[16:14:48] Eugene: I think you should take some typing lessons. Or just type more accurately.

[16:15:04] Judge: That's very rude.

[16:15:09] Eugene: Ummm ... Frankly, I didn't get your question :-)

[16:15:49] Judge: "Ummm" isnt a word. i think you need typing lessons

[16:15:58] Eugene: My thoughts are same. By the way, I still don't know your specialty – or, possibly, I've missed it?

# Εφαρμογές Επεξεργασίας Φυσικής Γλώσσας

---

- αυτόματη μετάφραση κειμένων
- έξυπνη αναζήτηση στο διαδίκτυο
- ανάκτηση πληροφορίας από σώματα κειμένων – δημιουργία περίληψης κειμένου
- κατηγοριοποίηση κειμένων
- κατανόηση κειμένου από τον υπολογιστή
- σύνθεση φυσικής γλώσσας
- δημιουργία φιλικών προς το χρήστη και εύχρηστων διεπαφών
- υποστήριξη συγγραφέα (διόρθωση ορθογραφικών και συντακτικών λαθών)
- Web Sentiment Analysis

# Εφαρμογές Επεξεργασίας Ομιλίας

---

- Αυτοματοποιημένες υπηρεσίες μέσω τηλεφώνου
- Φωνητικός έλεγχος μηχανών/συσκευών
- Συστήματα υπαγόρευσης
- Συστήματα αναγνώρισης ομιλητή

# Παράδειγμα: Αυτόματη μετάφραση

SYSTRAN Language Translation Technology - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.systran.com/index.html>

**SYSTRAN**  
Language Translation Technologies

Company | Products & Services | Solutions | Support | Contact | Purchase

Quick Links  
Technology  
Investors  
Press Releases  
Past Events  
GTC - Sacramento 2006  
Careers  
News  
August 04, 2006  
SYSTRAN announced its First Semester Interim Report [PDF]  
June 6, 2006  
Treasury Stock Acquisition Plan 2006 [PDF]  
Contact  
Authorized Resellers  
Affiliate Program

Translate on your PDA  
anytime,  
anywhere!

Solutions for  
Individual Users & Home Office  
Small & Medium Businesses  
Large Businesses  
Government & Public Administrations

What product do you need?  
Use our Product Comparison to find the right solution.  
Start Here

Already know what you need?  
Choose a product...

Translate with SYSTRAN 5.0  
Texts: Enter up to 150 words for translation  
Type your text here.  
Web page: Enter the Web address of the page you wish to translate  
http://www.  
English to French  
English to Arabic  
English to Simplified Chinese  
English to Traditional Chinese  
English to Dutch  
English to French  
English to German  
English to Italian  
English to Japanese  
English to Korean  
English to Portuguese  
English to Russian  
Click here to translate

Buy now from \$179.00

Need help?  
Access our Support Pages.  
Click Here

regorgization of errors, and effective dictionary  
significant cost and time savings, as well as improved  
consistency in translations.

Start 2 Microsoft ... 2 Internet ... 2 Internet ... Microsoft Po... Polylexicon Unribled - Paint 5:41 μμ

## Translated text in German

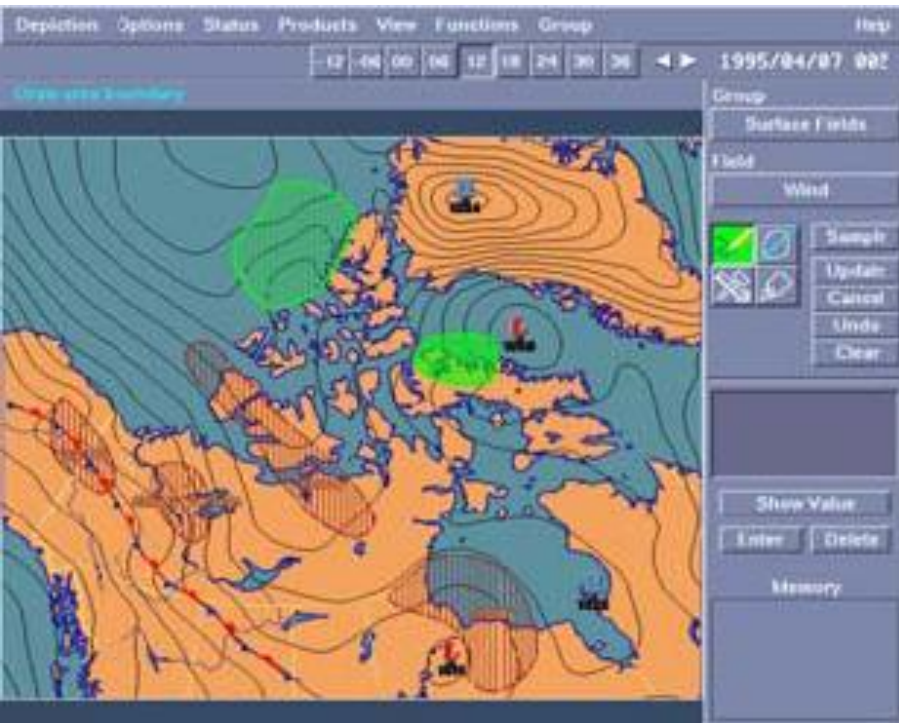
Computer werden jetzt als Begleiter, nicht nur als Mittel für das Erhöhen von Produktion gesehen.

## Translate with SYSTRAN 5.0

Text: Enter up to 150 words for translation

Computers are now seen as companions, not only as means for enhancing production.

# Παράδειγμα: Σύνθεση



είσοδος



έξοδος

Goldberg E: FoG: Synthesizing forecast text directly from weather maps

# Παράδειγμα: Εξαγωγή πληροφορίας

---

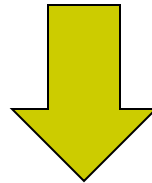
19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

- **Incident type** bombing
- **Date** March 19
- **Location** El Salvador: San Salvador (city)
- **Perpetrator** urban guerilla commandos
- **Physical target** power tower
- **Human target** -
- **Effect on physical target** destroyed
- **Effect on human target** no injury or death
- **Instrument** bomb

# Web Sentiment Analysis

---

I love you but hate the current political climate.



[no positive emotion or energy] 1-2-3-4-5 [very strong positive emotion]  
[no negative emotion] 1-2-3-4-5 [very strong negative emotion]

# Διεπιστημονική περιοχή

---

- Γλωσσολογία: κατανόηση της χρήσης της γλώσσας
- Ψυχολογία: κατανόηση της αντίληψης και της παραγωγής της γλώσσας
- Επιστήμη Υπολογιστών: ανάπτυξη καλύτερων και πιο εύχρηστων εφαρμογών
- Τεχνητή Νοημοσύνη: ανάπτυξη μοντέλων ανθρώπινου συλλογισμού



# Όνομα Επιστημονικής Περιοχής

---

- Επεξεργασία/Τεχνολογία Φυσικής Γλώσσας  
(Natural Language Processing)
  - Πληροφορική, Επιστήμη Υπολογιστών
- Υπολογιστική Γλωσσολογία  
(Computational Linguistics)
  - Γλωσσολογία
- Υπολογιστική Ψυχογλωσσολογία  
(Computational psycho-linguistics)
  - Ψυχολογία

# Eliza (<http://www-ai.ijs.si/eliza/eliza.html>)

---

- Βασίζεται στην ψυχιατρική προσέγγιση του Carl Rogers (~1940-1950). Ο θεραπευτής δεν προτείνει, δεν συμβουλεύει, αλλά προτρέπει τον ασθενή να συνεχίσει να μιλάει, ώστε να καταλήξει μόνος του στην λύση του προβλήματός του.
- Η Eliza είναι ένα σύστημα που αναγνωρίζει πρότυπα λέξεων/φράσεων μέσα στην ερώτηση του χρήστη και αποκρίνεται αντίστοιχα.

U: You don't argue with me.

S: Why do you think I don't argue with you?

U: I'm feeling a bit depressed

S: Are you depressed often?

U: I'm feeling a bit jump XyZZy

S: Why do you say you are feeling a bit jump XyZZy?

# Η γνώση στην γλώσσα

---

- Τι είναι οι λέξεις (ουσιαστικά, ρήματα, άρθρα);
- Πως συνδέονται μεταξύ τους;
- Τι σημαίνουν οι λέξεις;
- Πως συμβάλλουν τα νοήματα των λέξεων στις προτάσεις;

# Επίπεδα γλωσσολογικής γνώσης

---

- ❑ **Φωνητική και φωνολογική γνώση:**  
Πως σχηματίζονται οι λέξεις από ήχους.
- ❑ **Μορφολογική γνώση:**  
Πως σχηματίζονται οι λέξεις από μορφήματα: *έ-τρεχ-α*
- ❑ **Συντακτική γνώση:**  
Πως σχηματίζονται οι προτάσεις από λέξεις.
- ❑ **Σημασιολογική γνώση:**  
Το νόημα των λέξεων και (συνδυαζόμενα) νόημα προτάσεων.
- ❑ **Πραγματολογική γνώση:**  
Πως οι προτάσεις χρησιμοποιούνται σε διαφορετικά συμφραζόμενα, και πως τα συμφραζόμενα, και η γνώση του κόσμου επηρεάζουν την ερμηνεία μιας λέξης.

# Παράδειγμα: «*Αυτός αγαπάει τη Μαίρη.*»

---

## □ Φωνολογία

/a/ /f/ /t/ /o/ /s/ /a/ /g/ /a/ /p/ /a/ /i/ /t/ /i/ /m/ /e/ /r/ /i/

## □ Μορφολογία

**αυτ-ός:** αντωνυμία, ενικός, ονομαστική, αρσενικό

**αγαπ-άει:** ρήμα, ενικός, τρίτο πρόσωπο, ενεστώτας

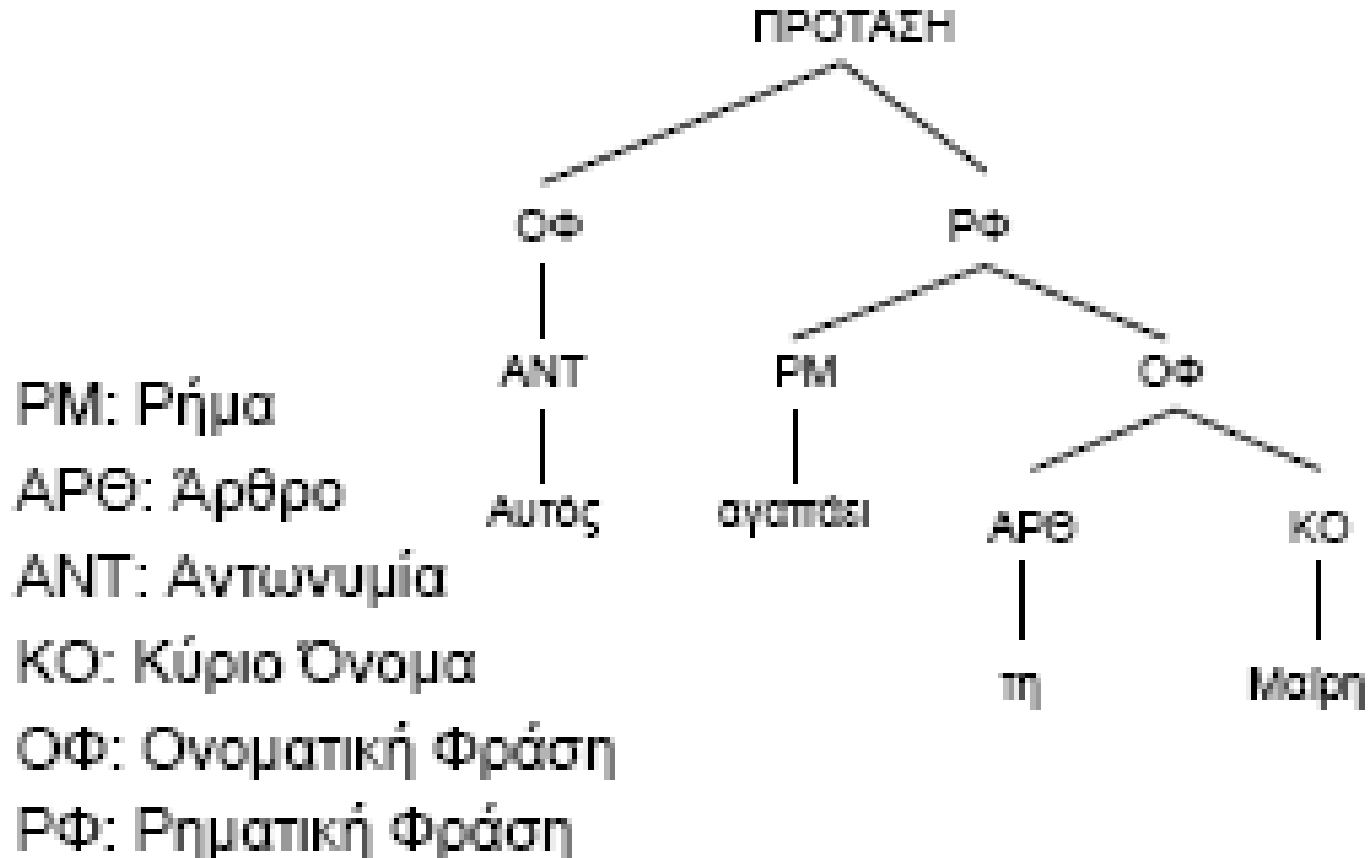
**τη:** οριστικό άρθρο, ενικός, θηλυκό

**Μαίρη:** κύριο όνομα, θηλυκό

**. :** σημείο στίξης, τέλος περιόδου

# Παράδειγμα: «Αυτός αγαπάει τη Μαίρη.»

## □ Σύνταξη



# Παράδειγμα: «*Αυτός αγαπάει τη Μαίρη.*»

---

- Σημασιολογία

$\exists x \text{ love}(x, \text{mary})$

- Πραγματολογία

$\text{love}(\text{john}, \text{mary})$

# Άλλα παραδείγματα

---

«- *Οι πράσινοι βάτραχοι έχουν μεγάλες μύτες*»

- **Σωστή:** Σύνταξη, Σημασιολογία
- **Λάθος:** Πραγματολογία

«- *Οι πράσινες ιδέες έχουν μεγάλες μύτες*»

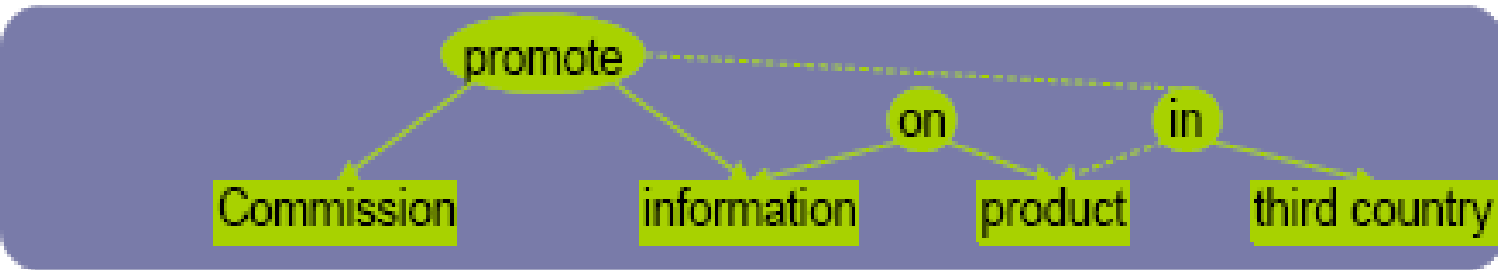
- **Σωστή:** Σύνταξη,
- **Λάθος:** Σημασιολογία, Πραγματολογία

«- *Μεγάλες έχουν οι πράσινες ιδέες μύτες*»

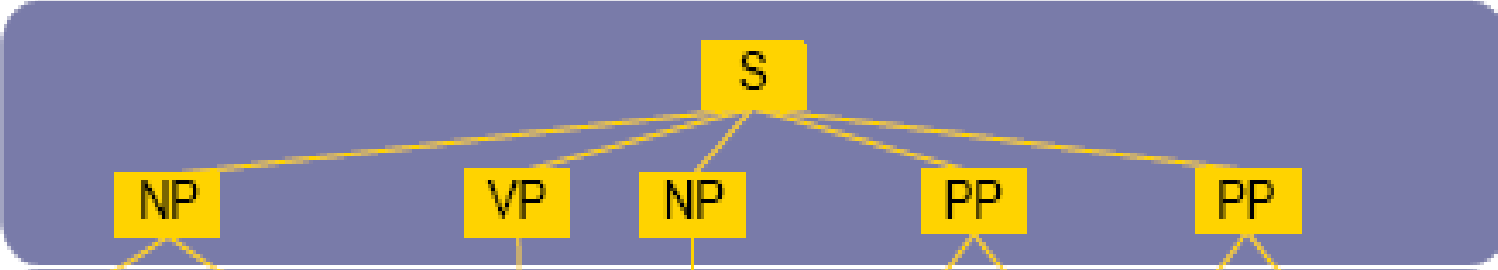
- **Λάθος:** Σύνταξη, Σημασιολογία, Πραγματολογία



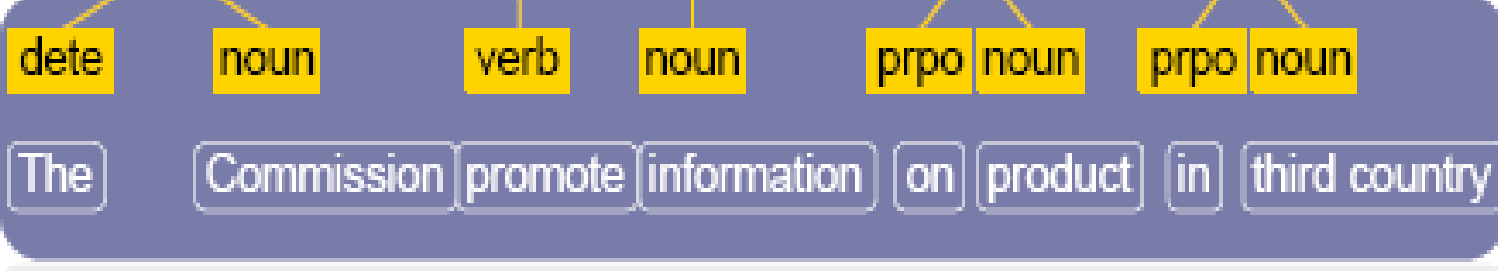
Sem



Syn



Lex



Text

The Commission promotes information on products in third countries

# Οι δυσκολίες της γλώσσας

---

- Ασάφεια (ambiguity)
  - Φωνολογική ([filo])
  - Μορφολογική (να δώσεις – οι δώσεις)
  - Συντακτική (Είδα τον άντρα με το τηλεσκόπιο)
  - Σημασιολογική (Έπιασα το ποντίκι)
  - Πραγματολογική (Τι μέρα έχουμε;)

Κάθε είδος ασάφειας απαιτεί γνώση παραπάνω επιπέδων για να επιλυθεί.

# Οι δυσκολίες της γλώσσας (συν.)

---

- Διαφορετικοί τρόποι έκφρασης
  - *«Πότε γεννήθηκε η Μαίρη;»*
  - *«Πότε είναι τα γενέθλια της Μαίρης;»*
- Πληροφορία που εννοείται
  - «Τηλεφώνησα στο σπίτι της Λένας για να πάμε σινεμά. Συμφώνησε με χαρά.»*
  - Η Λένα ήταν στο σπίτι της
  - Η Λένα απάντησε στο τηλέφωνο
  - Της ζήτησα να πάμε σινεμά
- Η ίδια πρόταση μπορεί να έχει διαφορετικό νόημα σε διαφορετικές περιπτώσεις
  - «Πώς είναι το νερό;»*
- Συνεχόμενη δημιουργία νέων λέξεων/εκφράσεων
  - «Θα το σκανάρω.»*

# Συνήθειες εργασίες στην ΕΦΓ

---

- Επισημείωση μερών του λόγου (part-of-speech tagging)
- Ανίχνευση ορίων φράσεων μέσα σε μια πρόταση (phrase boundary detection)
- Συντακτική ανάλυση (parsing)
- Άρση αμφισημίας λέξεων (word sense disambiguation)
- Αναγνώριση ονομάτων-οντοτήτων (named-entity recognition)
- Ανάθεση σημασιολογικών ρόλων (semantic role assignment)

# Έρευνα αιχμής

---

- Εξαγωγή πληροφορίας από κείμενα συγκεκριμένης θεματικής περιοχής
  - Οικονομικά
  - Βιο-ιατρική
- Αυτόματη περίληψη μεγάλων κειμένων
- Αυτόματη μετάφραση προφορικού λόγου
- Ανάπτυξη πρακτόρων με ανθρώπινα χαρακτηριστικά συνομιλίας
- ...

# Γλώσσες προγραμματισμού

---

- Prolog & Lisp
  - Εύκολος χειρισμός λιστών
  - Πρωτότυπα εφαρμογών τεχνητής νοημοσύνης
- Perl
  - Πληθώρα έτοιμων συναρτήσεων χειρισμού αλφαριθμητικών
- C & Java
  - Εμπορικές και web-based εφαρμογές
  - Ευρεία φορητότητα

# Ιστορικά στοιχεία

---

- Ορθολογική προσέγγιση (1965-1985)
  - Βασίζεται στην θεωρία του γλωσσολόγου Noam Chomsky
  - Προσπάθεια κωδικοποίησης (θεωρητικής μοντελοποίησης) της γλωσσολογικής γνώσης
  - Η γλώσσα θεωρείται ως μια ακολουθία συμβόλων

# Ιστορικά στοιχεία (συν)

---

- Εμπειρική/στοχαστική προσέγγιση (1950-1965 και 1985-σήμερα)
  - Βασίζεται στη θεωρία του Claude Shannon για την απόδοση πιθανοτήτων σε γλωσσολογικές δομές
  - Η σύνθετη και περίπλοκη δομή της γλώσσας μαθαίνεται καθορίζοντας ένα γενικό μοντέλο της γλώσσας και λαμβάνοντας τις τιμές των παραμέτρων μέσω στατιστικών μεθόδων, αναγνώρισης προτύπων και μηχανικής μάθησης
  - Πρακτικά συστήματα βάσει σωμάτων κειμένων (corpus-based)



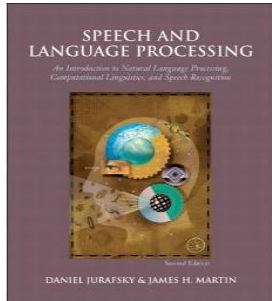
# Ιστορικά στοιχεία (συν)

---

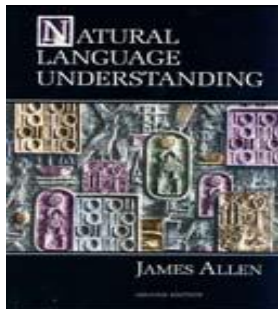
- 1950 – 1965: Έμφαση στη μηχανική μετάφραση
- 1966 – 1978: Έμφαση στη σημασιολογία
- 1979 – 1985: Έμφαση στη σύνταξη
- 1986 – σήμερα: Έμφαση στις εργασίες χαμηλού επιπέδου, αξιολόγηση

# Βιβλία

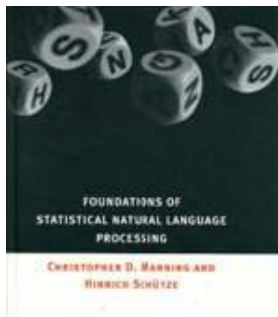
---



Jurafsky, D. and J. H. Martin:  
*Speech and Language Processing*  
Prentice-Hall. 2008.  
ISBN 0131873210



Allen, J.:  
*Natural Language Understanding* (2<sup>nd</sup> edition)  
The Benajmins/Cummings Publishing Company Inc. 1994.  
ISBN 0-8053-0334-0



Manning, C. D. and H. Schütze:  
*Foundations of Statistical Natural Language Processing*  
The MIT Press. 1999.  
ISBN 0-262-13360-1

# Διευθύνσεις

---

- Association for Computational Linguistics  
(<http://www.aclweb.org>)
- Ινστιτούτο Επεξεργασίας του Λόγου  
(<http://www.ilsp.gr>)
- The Natural Language Software Registry  
(<http://registry.dfki.de>)
- The Natural Language Toolkit  
(<http://nltk.sourceforge.net>)
- Survey of the State of the Art in Human Language Technology  
(<http://www.cse.ogi.edu/CSLU/HLTsurvey/>)