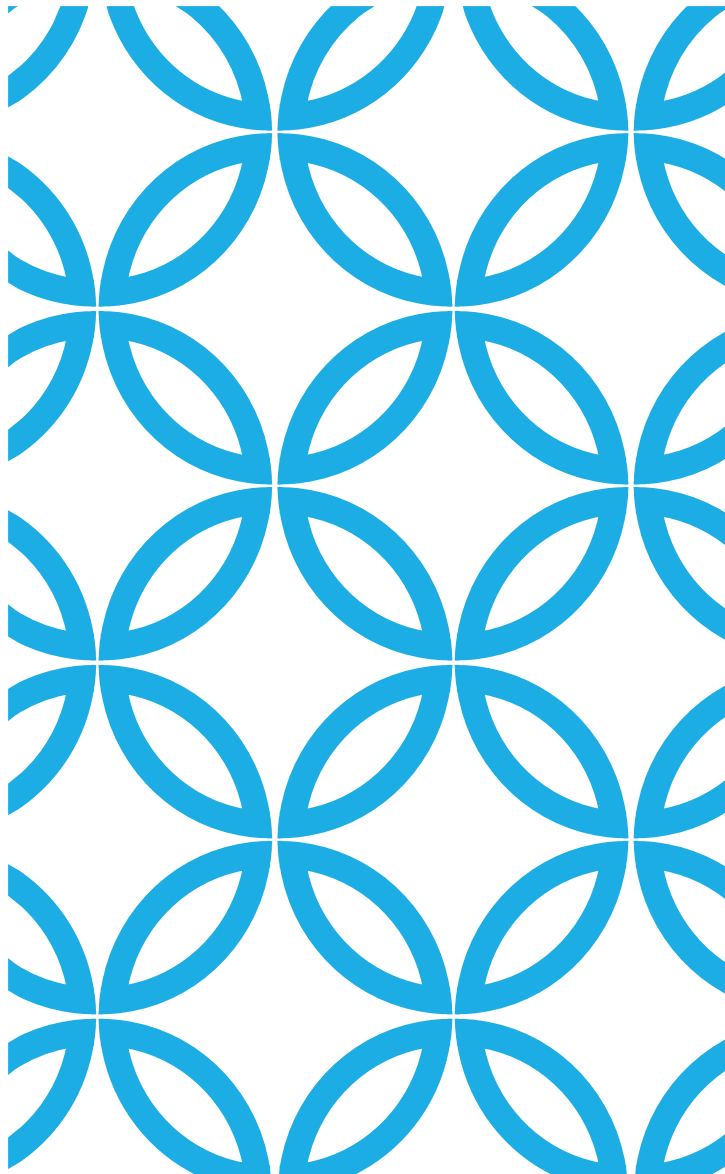




ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Κάτια Κερμανίδου
kerman@ionio.gr

Εισαγωγή στη Γλωσσική Τεχνολογία

Κάποιες εικόνες είναι δανεισμένες από την Diyi Yang στο Georgia Institute of Technology και από τους Sun & Chen στο University of Cambridge

Γλωσσική Τεχνολογία

- Η ερευνητική περιοχή που ασχολείται με την σχεδίαση, ανάπτυξη και αξιολόγηση εργαλείων, τεχνολογιών και μεθοδολογιών για την επεξεργασία της φυσικής (ανθρώπινης) γλώσσας
- Τι είναι γλώσσα;
 - Δεν είναι προφανές. Είναι τα emojis κομμάτι της γλώσσας;
- Οι μορφές του Λόγου
 - Γραπτός λόγος – Επεξεργασία Φυσικής Γλώσσας
 - Προφορικός λόγος – Επεξεργασία Ομιλίας
 - Χειρόγραφο κείμενο - Οπτική Αναγνώριση Χαρακτήρων

Φυσικές
γλώσσες

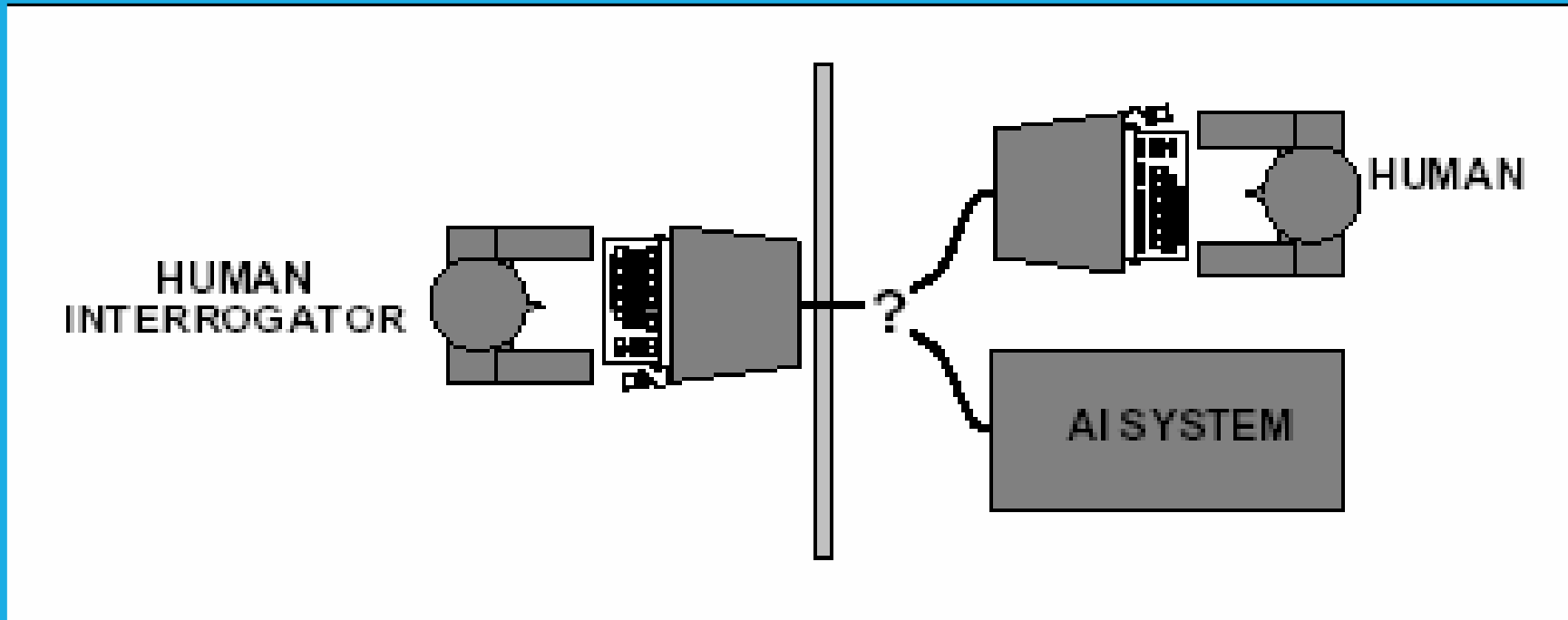
C, Lisp, Python, Ruby, Scala, ...
English, Welsh, Afrikaans, Mandarin, ...
English as a Second Language, ...
sign languages, ...
Sanskrit, ...
dolphin language
...

<https://www.cl.cam.ac.uk/teaching/2425/L390/doc/>

Συναφείς επιστημονικές περιοχές

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

Test Turing



Εφαρμογές Γλωσσικής Τεχνολογίας

NLP Applications

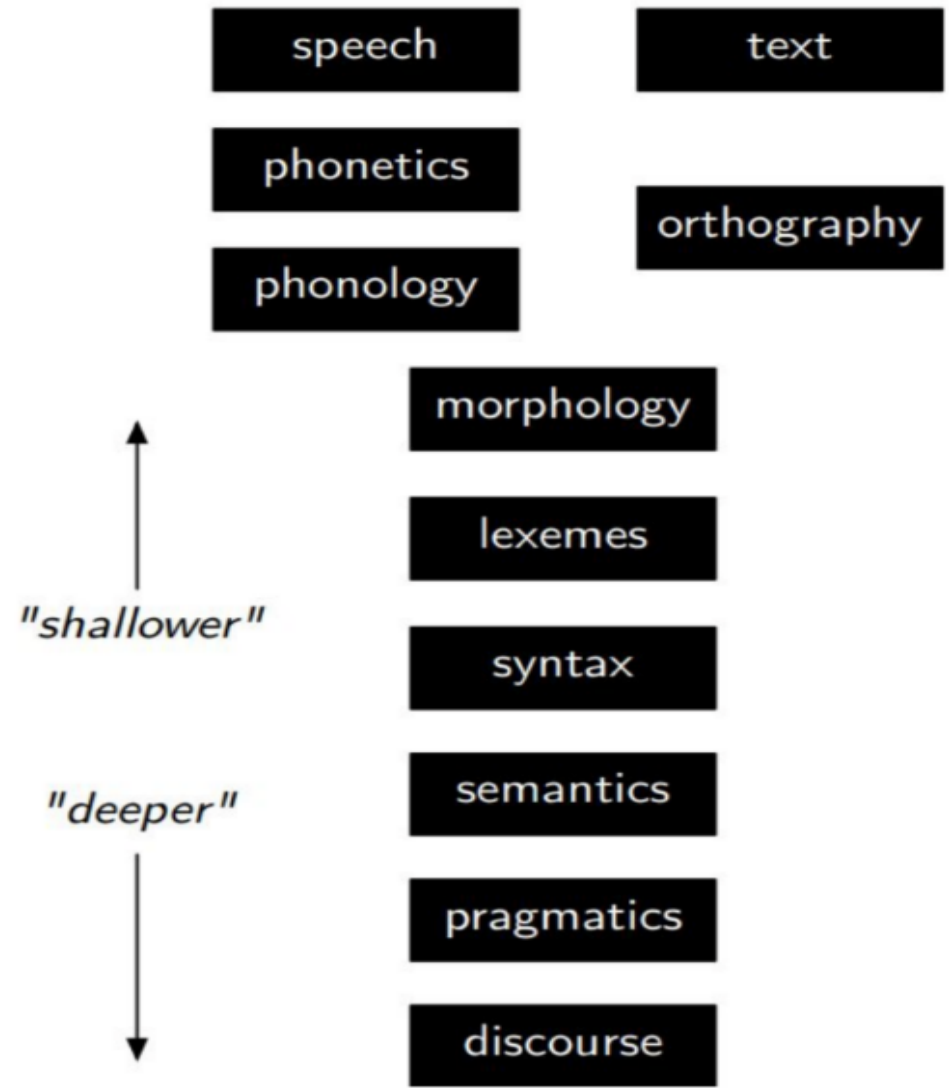
- Αυτόματη μετάφραση
- Έξυπνη αναζήτηση στο διαδίκτυο
- Ανάκτηση πληροφορίας από σώματα κειμένων (Information Retrieval)
- Εξαγωγή πληροφορίας από σώματα κειμένων (Information Extraction)
- Δημιουργία περίληψης κειμένου
- Κατηγοριοποίηση κειμένων
- Κατανόηση κειμένου από τον υπολογιστή
- Σύνθεση φυσικής γλώσσας
- Δημιουργία φιλικών προς το χρήστη και εύχρηστων διεπαφών
- Question answering
- Web Sentiment Analysis
- Διαλογικά Συστήματα
- Επεξεργασία Ομιλίας
- Σύνθεση Ομιλίας (TTS)

Εργασίες Γλωσσικής Τεχνολογίας

NLP Tasks/Technologies

- Αναγνώριση Μερών του Λόγου (Part of Speech Tagging)
- Αναγνώριση οντοτήτων (Named Entity Recognition)
- Αναγνώριση Φράσεων (Chunking)
- Συντακτική Ανάλυση (Parsing)
- Σημασιολογική Ανάλυση (Semantic Processing)
- Αναγνώριση Σημασιολογικών ρόλων (Semantic Role Labeling)
- Άρση Αμφισημίας Έννοιας Λέξης (Word Sense Disambiguation)
- Μοντελοποίηση Γλώσσας (Language Modeling)
- Άρση Αναφορικής Ασάφειας (Reference Resolution)
- Ανάλυση Λόγου (Discourse Analysis)

Level Of Linguistic Knowledge



Φωνητική/Φωνολογία Phonetics/Phonology

- Προφορά
- Φωνήματα (Ηχοι)

SOUNDS

Th

i

a

si

e

n

https://sites.cc.gatech.edu/classes/AY2020/cs7650_spring/slides/1_intro.pdf

Λέξεις

- Μοντελοποίηση Γλώσσας
- Αναγνώριση λεκτικών συμβόλων (tokenization)

WORDS

This is a simple sentence

https://sites.cc.gatech.edu/classes/AY2020/cs7650_spring/slides/1_intro.pdf

Μορφολογία

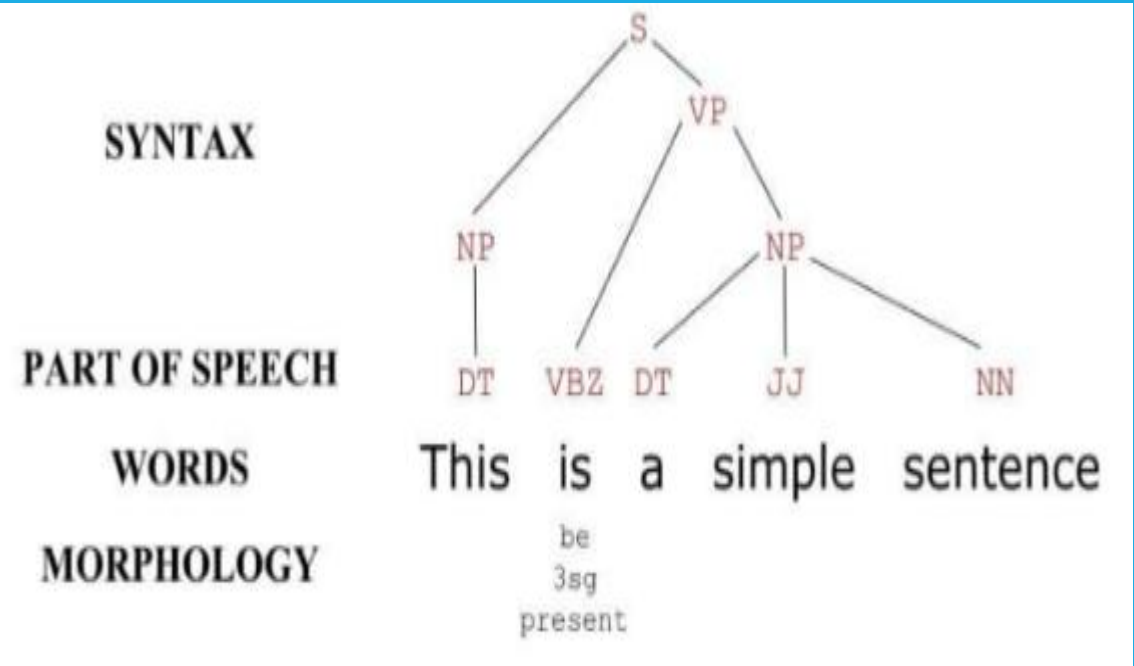
- Μορφολογική Ανάλυση
- Λημματοποίηση
- Ορθογραφικός έλεγχος
- Αναγνώριση κατάληξης (Stemming)
- Αναγνώριση Μερών του Λόγου

PART OF SPEECH	DT	VBZ	DT	JJ	NN
WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

https://sites.cc.gatech.edu/classes/AY2020/cs7650_spring/slides/1_intro.pdf

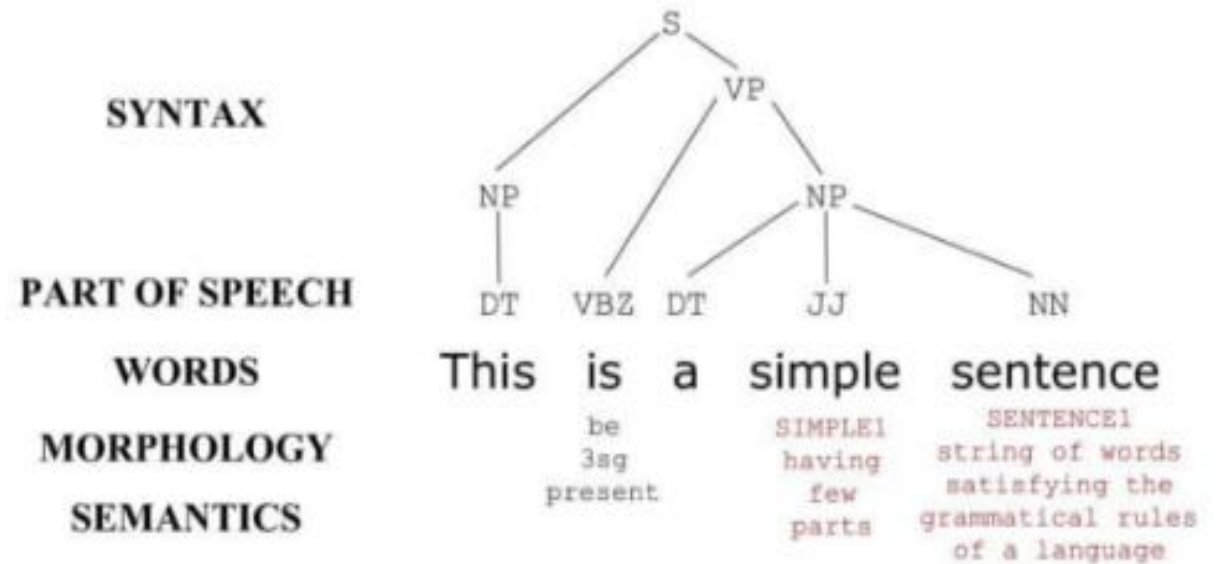
Σύνταξη

- Συντακτική ανάλυση
- Αναγνώριση συντακτικών δομών
- Αναγνώριση συντακτικών ρόλων
- Αναγνώριση συντακτικών εξαρτήσεων



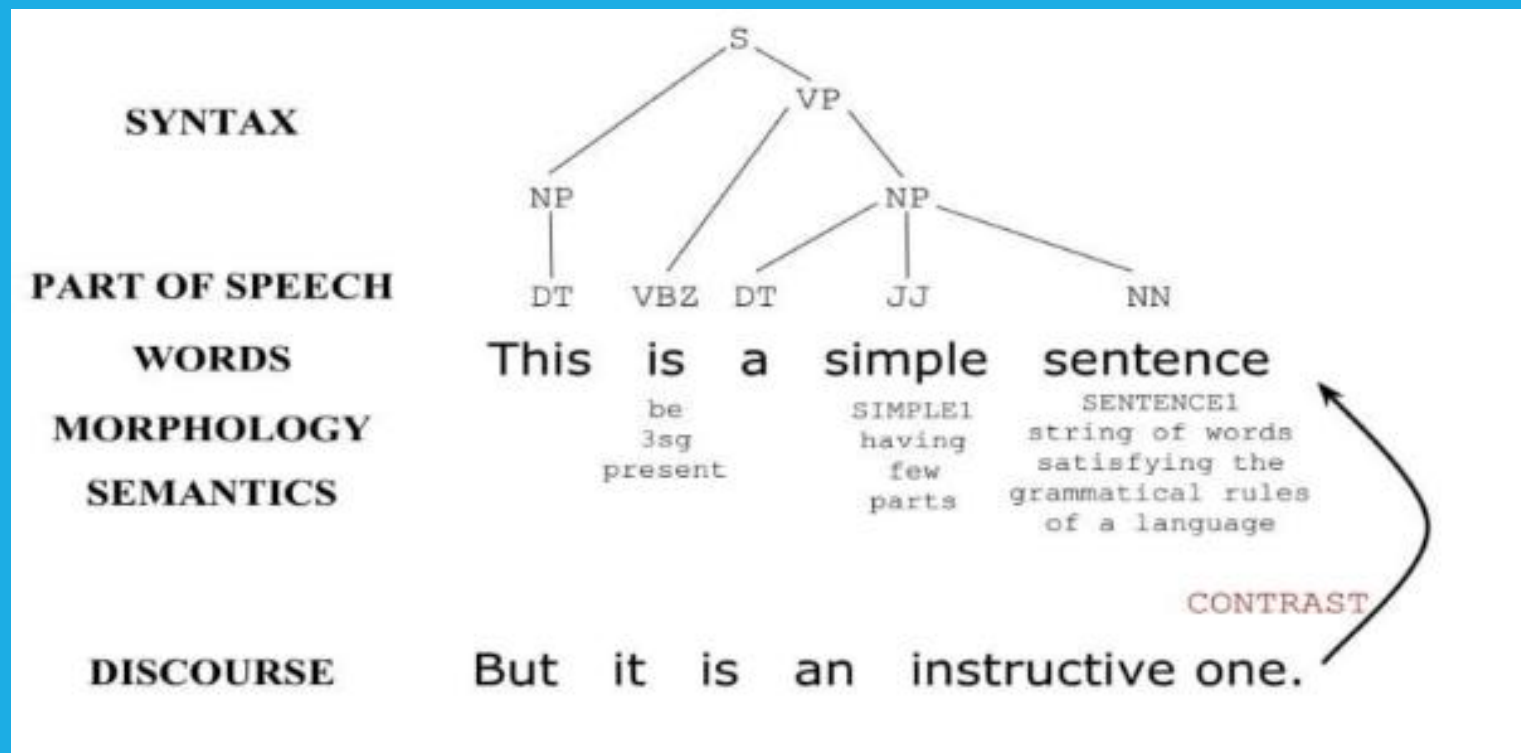
Σημασιολογία

- Σημασιολογική ανάλυση
- Σημασιολογική αναπαράσταση
- Αναπαράσταση γνώσης
 - πχ Κατηγορηματική Λογική
- Αναγνώριση σημασιολογικών ρόλων
- Άρση αμφισημίας έννοιας λέξης



Πραγματολογία/Λόγος

- Διαλογικά Συστήματα
- Αναφορική Ασάφεια
- Συνοχή Λόγου
- Ελλειπτικές Προτάσεις



NLP tasks - CoNLL shared task

input words	
1	U.N.
2	official
3	Ekeus
4	heads
5	for
6	Baghdad
7	.

part-of-speech tags	
1	U.N.
2	official
3	Ekeus
4	heads
5	for
6	Baghdad
7	.

input words		part-of-speech tags	
1	U.N.	NNP	B-NP
2	official	NN	I-NP
3	Ekeus	NNP	I-NP
4	heads	VBZ	B-VP
5	for	IN	B-PP
6	Baghdad	NNP	B-NP
7	.	.	0

input words		part-of-speech tags		syntactic chunks	named entities
1	U.N.	NNP	B-NP	B-ORG	
2	official	NN	I-NP	0	
3	Ekeus	NNP	I-NP	B-PER	
4	heads	VBZ	B-VP	0	
5	for	IN	B-PP	0	
6	Baghdad	NNP	B-NP	B-LOC	
7	.	.	0	0	

input words		part-of-speech tags		syntactic dependencies	
1	U.N.	NNP	B-NP	B-ORG	3
2	official	NN	I-NP	0	3
3	Ekeus	NNP	I-NP	B-PER	4
4	heads	VBZ	B-VP	0	0
5	for	IN	B-PP	0	6
6	Baghdad	NNP	B-NP	B-LOC	4
7	.	.	0	0	4

<https://www.cl.cam.ac.uk/teaching/2425/L390/doc/>

Δυσκολίες/Προκλήσεις στην Γλωσσική Τεχνολογία

■ **Αμφισημία (Ambiguity)**

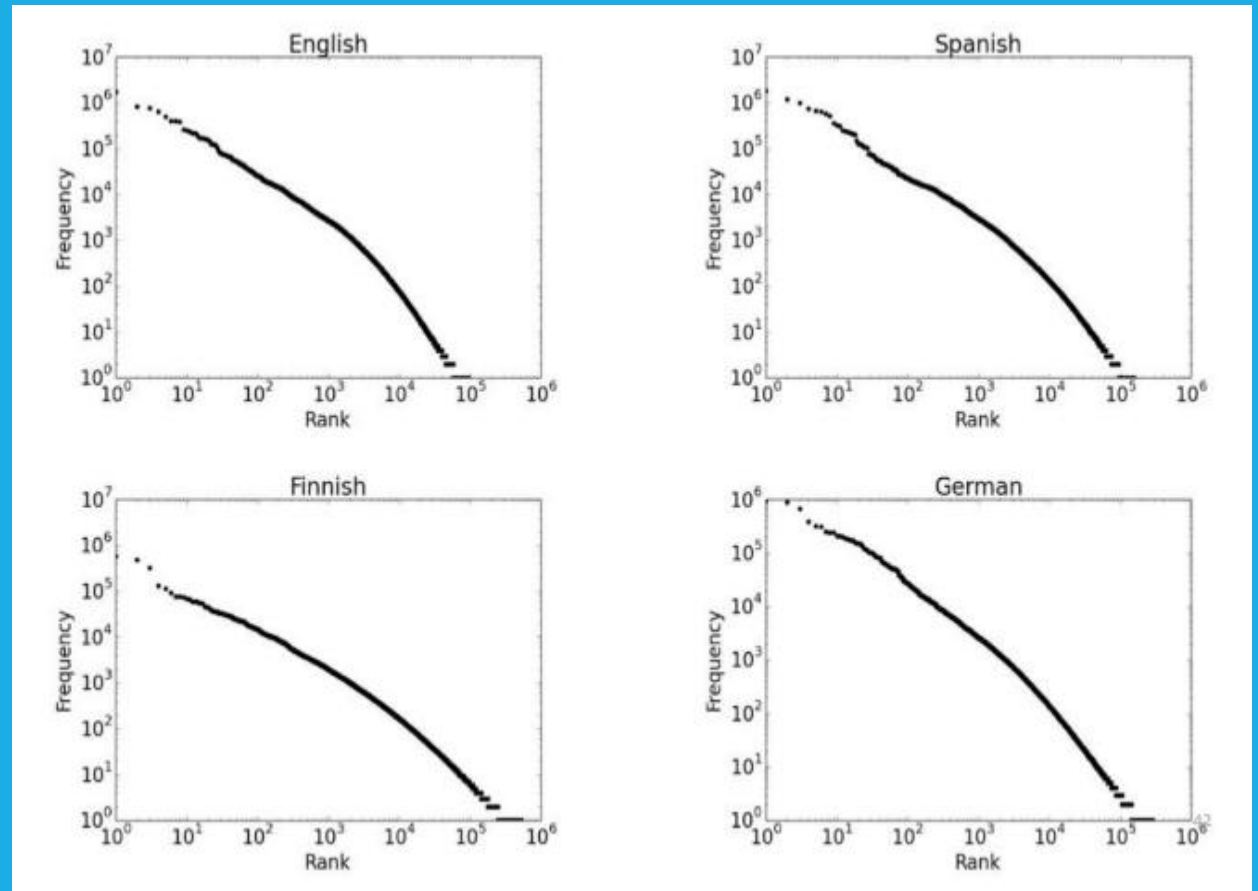
- Φωνολογική ([fɪlo])
- Μορφολογική (**Κούμπωσε** το παντελόνι.)
- Συντακτική (Είδα τον άντρα **με το τηλεσκόπιο**)
- Σημασιολογική (Έπιασα το **ποντίκι**)
- Πραγματολογική (Ο Γιάννης αγαπάει τον Πέτρο. Είναι καλός φίλος. **Ποιός;**)

■ **Πλεονασμός:** Το ίδιο νόημα μπορεί να εκφραστεί με πολλούς διαφορετικούς τρόπους

- She gave the pen to Anna / She gave Anna the pen
- Some friends stopped by / Some friends visited / A few friends paid a visit

Δυσκολίες/Προκλήσεις στην Γλωσσική Τεχνολογία

- **Σπανιότητα (Sparsity)**
- Νόμος του Zipf: Αν κατατάξουμε με φθίνουσα σειρά συχνότητας της λέξης μιας γλώσσας, η θέση μιας λέξης στη σειρά αυτή (rank) είναι αντιστρόφως ανάλογη της συχνότητάς της (frequency)
- Όσο μεγάλο σώμα κειμένων και αν πάρω για να το επεξεργαστώ και να κάνω μια οποιαδήποτε εργασία, πάντα θα υπάρχουν φαινόμενα που θα εμφανίζονται πολύ σπάνια ή και καθόλου
- Αυτό σημαίνει ότι θα πρέπει να βρω έναν τρόπο να εκτιμήσω πιθανότητες για φαινόμενα σπάνια ή ανύπαρκτα

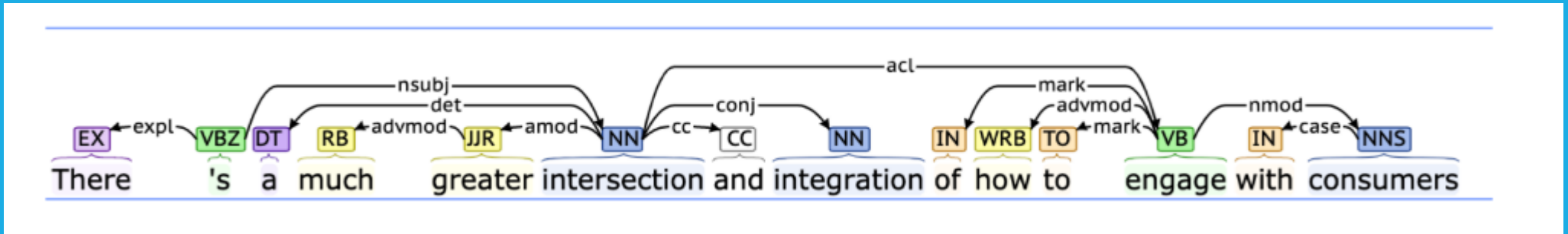


https://sites.cc.gatech.edu/classes/AY2020/cs7650_spring/slides/1_intro.pdf

Δυσκολίες/Προκλήσεις στην Γλωσσική Τεχνολογία

■ Διαφορετικότητα

- Η γλώσσα αλλάζει μορφές έκφρασης, ύφος, λεξιλόγιο, ανάλογα με το που χρησιμοποιείται. Έστω ένας συντακτικός αναλυτής που έχει εκπαιδευτεί στο σώμα κειμένων του Wall Street Journal.



https://sites.cc.gatech.edu/classes/AY2020/cs7650_spring/slides/1_intro.pdf

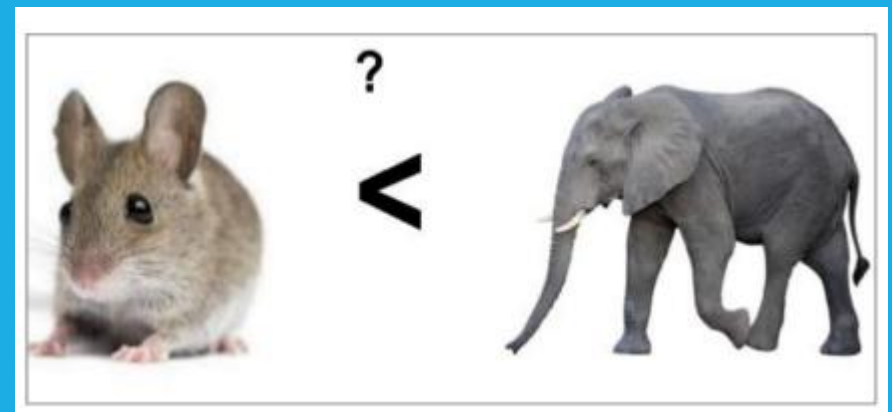
Πώς θα τα πάει αυτός ο αναλυτής σε ένα κείμενο κοινωνικού δικτύου;
Taking over @JB's Instagram today. Be on be the lookout for politics going pop! Good 4 u! lololol!!!

Twitter, Wikipedia, Πρακτικά Κοινοβουλίων/Συνελεύσεων, Επιστημονικά Άρθρα, Ομιλίες, Συνεντεύξεις, Τηλεφωνικές Συνομιλίες, Ειδησεογραφικά Κείμενα, Λογοτεχνία, Ποίηση, Διαφημιστικά κείμενα, Νομικά κείμενα

Δυσκολίες/Προκλήσεις στην Γλωσσική Τεχνολογία

■ Μοντελοποίηση

- Πώς μοντελοποιώ την έννοια μιας λέξης/πρότασης
- Πώς μοντελοποιώ την γνώση που έχω για τα συμφραζόμενα μέσα στα οποία εμφανίζεται μια λέξη/πρόταση;
- Πώς μοντελοποιώ την γνώση που έχω γενικά για τον κόσμο;



ΓΤ Βασισμένη σε Κανόνες / Βασισμένη στη Λογική (1965-1985)

- Η θεωρητική γλωσσολογική γνώση μοντελοποιείται με σύμβολα
- Η γλώσσα είναι μια ακολουθία συμβόλων
- Κανόνες χειρωνακτικοί αναπτύσσονται για την επεξεργασία της

Noam Chomsky

Article Talk

Read View source View history Tools

From Wikipedia, the free encyclopedia

"Chomsky" redirects here. For other uses, see [Chomsky \(disambiguation\)](#).

Avram Noam Chomsky^[a] (born December 7, 1928) is an American professor and [public intellectual](#) known for his work in [linguistics](#), political activism, and [social criticism](#). Sometimes called "the father of modern linguistics",^[b] Chomsky is also a major figure in [analytic philosophy](#) and one of the founders of the field of [cognitive science](#). He is a laureate professor of linguistics at the [University of Arizona](#) and an [institute professor](#) emeritus at the [Massachusetts Institute of Technology](#) (MIT). Among the most cited living authors, Chomsky has written more than 150 books on topics such as linguistics, war, and politics. In addition to his work in linguistics, since the 1960s Chomsky has been an influential voice on the [American left](#) as a consistent critic of [U.S. foreign policy](#), [contemporary capitalism](#), and [corporate influence](#) on political institutions and the media.

Noam Chomsky



Chomsky in 2017

Born Avram Noam Chomsky

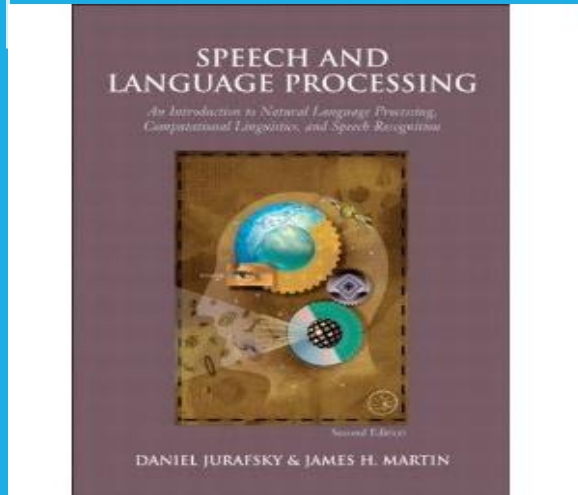
ΓΤ Καθοδηγούμενη από Δεδομένα / Πιθανοτική ΓΤ (1985-2015)

- Δεδομένα: Πραγματικά παραδείγματα χρήσης της γλώσσας = σώματα κειμένων
- Η γνώση για την γλώσσα εξάγεται αυτόματα με επεξεργασία σωμάτων κειμένων

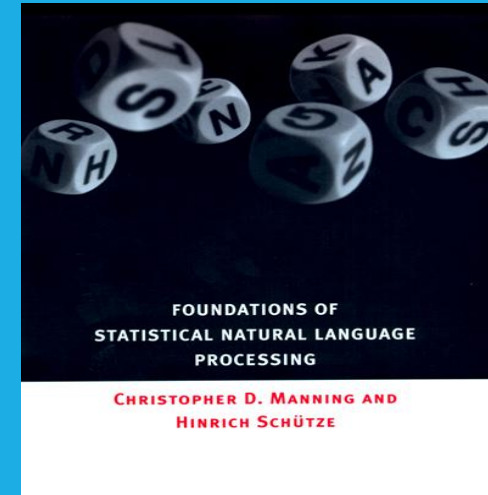
Συσχετιστική ΓΤ - Connectionist NLP (2015 – σήμερα)

- Η γνώση για την γλώσσα μοντελοποιείται μέσα από βαθιές διανυσματικές αναπαραστάσεις λέξεων/φράσεων/προτάσεων.
- Τα διανύσματα αυτά εξάγονται από βαθιά νευρωνικά δίκτυα τα οποία εκπαιδεύονται σε πολύ μεγάλα σώματα κειμένων
- Η οποιαδήποτε εργασία γλωσσικής τεχνολογίας πραγματοποιείται με εκπαίδευση αντίστοιχου νευρωνικού δικτύου

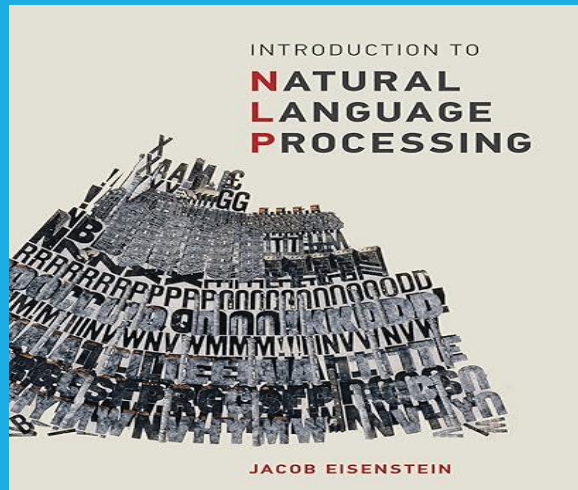
Βιβλία



<https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>



<https://nlp.stanford.edu/fsnlp/>



<https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf>