ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΕΛΟΠΟΝΝΗΣΟΥ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΓΛΩΣΣΟΛΟΓΙΑ

12η διάλεξη

Π. ΓΑΚΗΣ

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.

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;

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

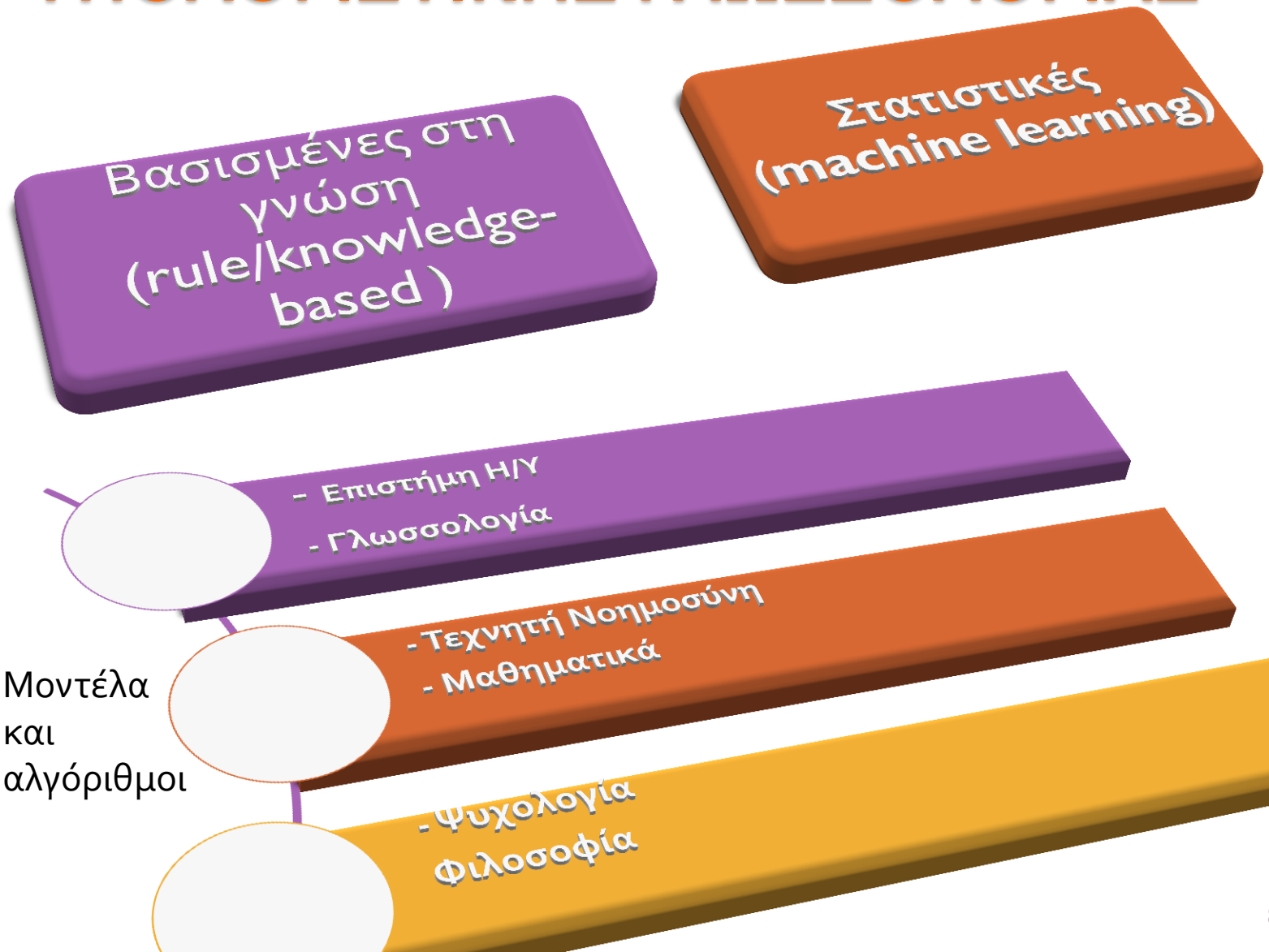
Επεξεργασία Φυσικής Γλώσσας (NLP)

- 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 (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

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

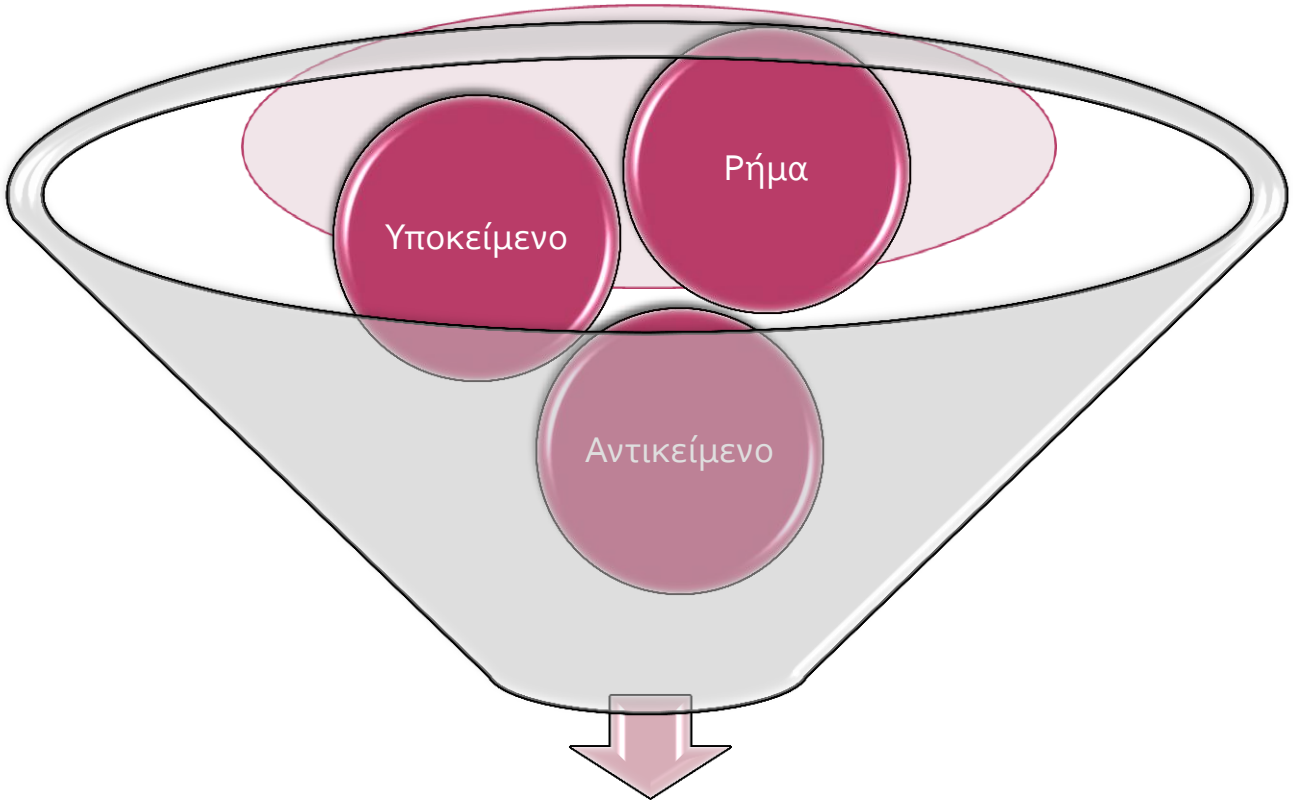
Ambiguity vs NLP

- 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.

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

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*»;

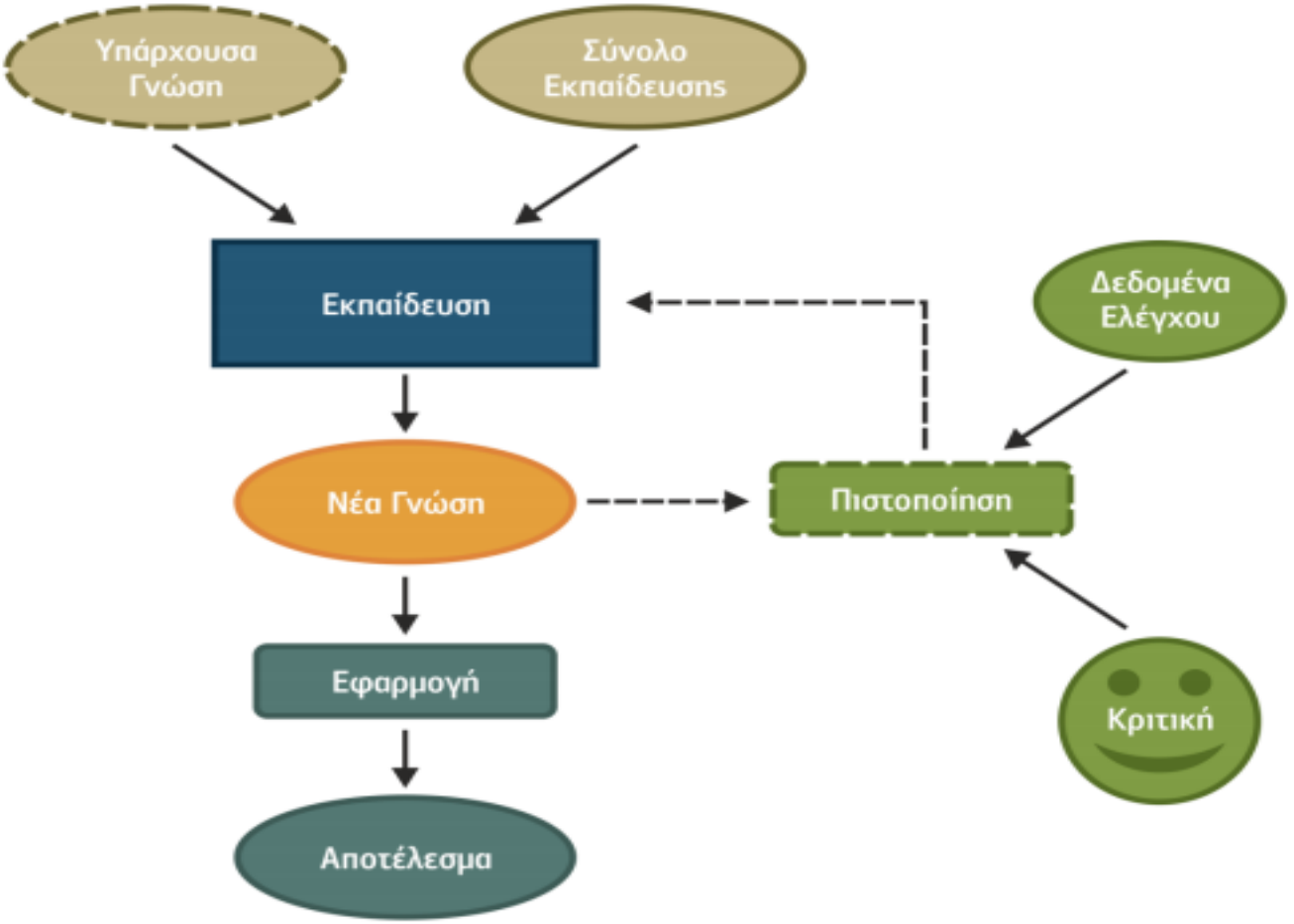
Machine Learning (μηχανική μάθηση)

- 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 (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.).»

ΥΠΟΛΟΓΙΣΤΙΚΑ ΛΕΞΙΚΑ

Υπολογιστικό Λεξικό

- φέρει πληροφορία: α) ορθογραφική (ορθή γραφή του κλιτικού τύπου),
- β) μορφηματική (το είδος των μορφημάτων: πρόθημα, θέμα, επίθημα, κατάληξη, που απαρτίζουν τον κλιτικό τύπο),
- γ) μορφοσυντακτική (μέρος του λόγου, γένος, πτώση, πρόσωπο κ.λπ.),
- δ) υφολογική (τα υφολογικά χαρακτηριστικά του τύπου: προφορικό, λόγιο κ.λπ.) και
- ε) ορολογική (επιπλέον πληροφορία για το αν ο τύπος αποτελεί μέρος ειδικού λεξιλογίου)

Υπολογιστική Λεξικογραφία

Κέρδη:

- Γνώση σχετικά με **μεμονωμένες λέξεις μιας γλώσσας**: απαραίτητη για κάθε είδους επεξεργασία φυσικής γλώσσας
- **Συστήματα Μηχανικής Μετάφρασης**: από τις πρώτες εφαρμογές υπολογιστικής λεξικογραφίας Εφαρμογές Speech-to-Text και Text-to-Speech
- **Ηλεκτρονικά λεξικά τσέπης**: Εργαλεία ελέγχου ορθογραφίας
- **Κοινό αίτημα**: σχεδιασμός λεξικών πηγών ενιαίας μορφής έτσι ώστε να διευκολύνεται η ευρεία χρήση τους και να αποφεύγεται η επανάληψη των διαδικασιών ανάπτυξής τους

Υπολογιστικά Λεξικά

- Τα υπολογιστικά λεξικά είτε:

1. αποτελούν απαραίτητο στοιχείο για κάθε σύστημα επεξεργασίας φυσικής γλώσσας (ΕΦΓ) είτε
2. είναι αναγνώσιμα από ανθρώπους, όπως οι λεξιλογικές βάσεις δεδομένων του Wordnet και του Framenet

Χαρακτηριστικά Υπολογιστικών λεξικών

- Χαρακτηριστικό στοιχείο των υπολογιστικών λεξικών είναι **ο απεριόριστος χώρος αποθήκευσης δεδομένων** όπως επίσης και **η δυνατότητα αποθήκευσης περισσότερων πληροφοριών σε περισσότερα επίπεδα**. *The absence of space constraints call for more, not less intellectual discipline in the selection and arrangement of information'*(Hanks, 2001).

Χαρακτηριστικά Υπολογιστικών λεξικών

- Υπάρχει επιπλέον **μεγαλύτερη ταχύτητα** κατά την ανάκληση, επεξεργασία και παρουσίαση των δεδομένων, **δυνατότητα ταυτόχρονης αναζήτησης σε πολλά λεξικά** ταυτόχρονα όπως επίσης και **δυνατότητα συχνότερης συντήρησης και επικαιροποίησης** σε σχέση με τα έντυπα.

ΟΝΟΜΑΤΙΚΕΣ ΟΝΤΟΤΗΤΣ

Μορφές των ΟΟ

- Μια ΟΟ μπορεί να παρουσιάζει ποικίλες μορφές.
- Ειδικότερα, οι ημερομηνίες εμφανίζονται είτε αριθμητικά (*14/12* ή *14-12*) είτε με γλωσσικά μέσα, π.χ. *δεκατέσσερις Δεκεμβρίου* είτε σε συνδυασμό αριθμητικής και λεκτικής μορφής (αλφαριθμητικά), π.χ. *14 Δεκεμβρίου 2013*.
- Επιπλέον, μια ΟΟ μπορεί να περιέχει εκτός από ελληνικούς και λατινικούς χαρακτήρες (*το Express Samina, το αεροπλάνο Airbus A316*).
- Στο συντακτικό επίπεδο, οι ΟΟ απαντούν στις ερωτήσεις *ποιος, πού, πότε*.

Δυσκολίες κατηγοριοποίησης ονοματικών οντοτήτων

- Η κατηγοριοποίηση των ΟΟ δεν είναι πάντα ευχερής, καθώς φαινόμενα όπως (α) η **ομωνυμία,** (β) η **πολυσημία** και (γ) η **μετωνυμία** δυσχεραίνουν τη διαδικασία αυτή:

1. Ο τύπος αρκτικολέξου *Α.Ε.* αποτελεί τη συντομευμένη μορφή δύο διαφορετικών ΟΟ, δηλ. ομώνυμων ΟΟ: α) *Ανώνυμη Εταιρεία* και β) *Ασφαλιστική Εταιρεία*.
2. Επίσης, εντάσσουμε στην κατηγορία των ομωνύμων και τις περιπτώσεις όπου ο ίδιος τύπος ΟΟ αναφέρεται σε διαφορετικά αντικείμενα αναφοράς, π.χ. ο τύπος *Ελευθέριος Βενιζέλος* μπορεί να αναφέρεται τόσο στον πολιτικό άνδρα, στο ιστορικό πρόσωπο, όσο και στον αερολιμένα της Αθήνας

Δυσκολίες κατηγοριοποίησης ονοματικών οντοτήτων

- Πολύσημες θεωρούμε τις ΟΟ όπου η ίδια μορφή αντιστοιχεί σε **πολλαπλές πλευρές** του ίδιου αντικειμένου αναφοράς. Λόγου χάρη, το ανθρωπωνύμιο *Αντώνης Σαμαράς* περιγράφει ταυτόχρονα ένα πολιτικό πρόσωπο, τον τ. πρωθυπουργό της Ελλάδας, όπως επίσης και τον τ. πρόεδρο του κόμματος *Νέα Δημοκρατία*. Στις περιπτώσεις αυτές στο λεξικό υπάρχει μία μόνο καταχώριση.
- Τέλος, η μετωνυμία είναι ένα σχήμα λόγου που μαζί με τη μεταφορά χρησιμοποιείται ευρύτατα στη γλώσσα για την κατονομασία νέων αντικειμένων, οντοτήτων κτλ. με ήδη υπάρχον υλικό. (π.χ. *11η Σεπτεμβρίου* εκτός από ημερομηνία αποτελεί ΟΟ η οποία μετωνυμικά περιγράφει ένα γεγονός:

Δυσκολίες εντοπισμού ονοματικών οντοτήτων

- Οι ΟΟ παρουσιάζουν δυσκολίες και στον εντοπισμό τους, καθώς συχνά απαντούν στα κείμενα υπό τη μορφή **αρκτικολέξου** ή **ακρωνυμίου:**

π.χ.

- *2α. … και κυρίως στις γραμματείες των περιφερειών και σε διοικήσεις ΔΕΚΟ*
- *2β. … απασχόλησε χθες το Ανώτατο Δικαστήριο, όπου προσέφυγε η διοίκηση των ΕΛΤΑ με αίτημα να αναιρεθεί προηγούμενη απόφαση.*

- η μορφή με την οποία εμφανίζονται οι ΟΟ δεν είναι πάντοτε σταθερή, ακόμα και μέσα στο ίδιο κείμενο

Δυσκολίες εντοπισμού ονοματικών οντοτήτων

- ζήτημα στον εντοπισμό των ΟΟ τίθεται, όπως είδαμε, από τα ομώνυμα αρκτικόλεξα

  π.χ. ο τύπος του αρκτικολέξου *Α.Ε.* αποτελεί τη συντομευμένη μορφή δύο διαφορετικών ΟΟ: α) *Ανώνυμη Εταιρεία* και β) *Ασφαλιστική Εταιρεία*.

- Η δυσκολία στην αυτόματη αναγνώρισή τους μπορεί να παραμείνει ακόμα και όταν ο ίδιος τύπος αρκτικολέξου αντιστοιχεί σε ΟΟ διαφορετικού γένους, όπως ο τύπος αρκτικολέξου *Γ.Σ.* που αντιστοιχεί στο α) *Γυμναστικός Σύλλογος*, ΟΟ **αρσενικού γένους**, και στο β) *Γενική Συνέλευση*, ΟΟ **θηλυκού γένους**.

Δυσκολίες εντοπισμού ονοματικών οντοτήτων

- Η δυσκολία εντοπισμού μιας ΟΟ κυρίως από συστήματα αυτόματης αναγνώρισης δομών σε κείμενα οφείλεται και στο γεγονός ότι η κατηγορία αυτή διαθέτει ανοιχτό, απροσδιόριστο και ετερογενή αριθμό λημμάτων, και επομένως καθίσταται **αδύνατη η πλήρης καταχώρισή τους σε λεξικά είτε έντυπα είτε ηλεκτρονικά**.

ΑΥΤΟΜΑΤΗ ΜΕΤΑΦΡΑΣΗ

Αυτόματη Μετάφραση

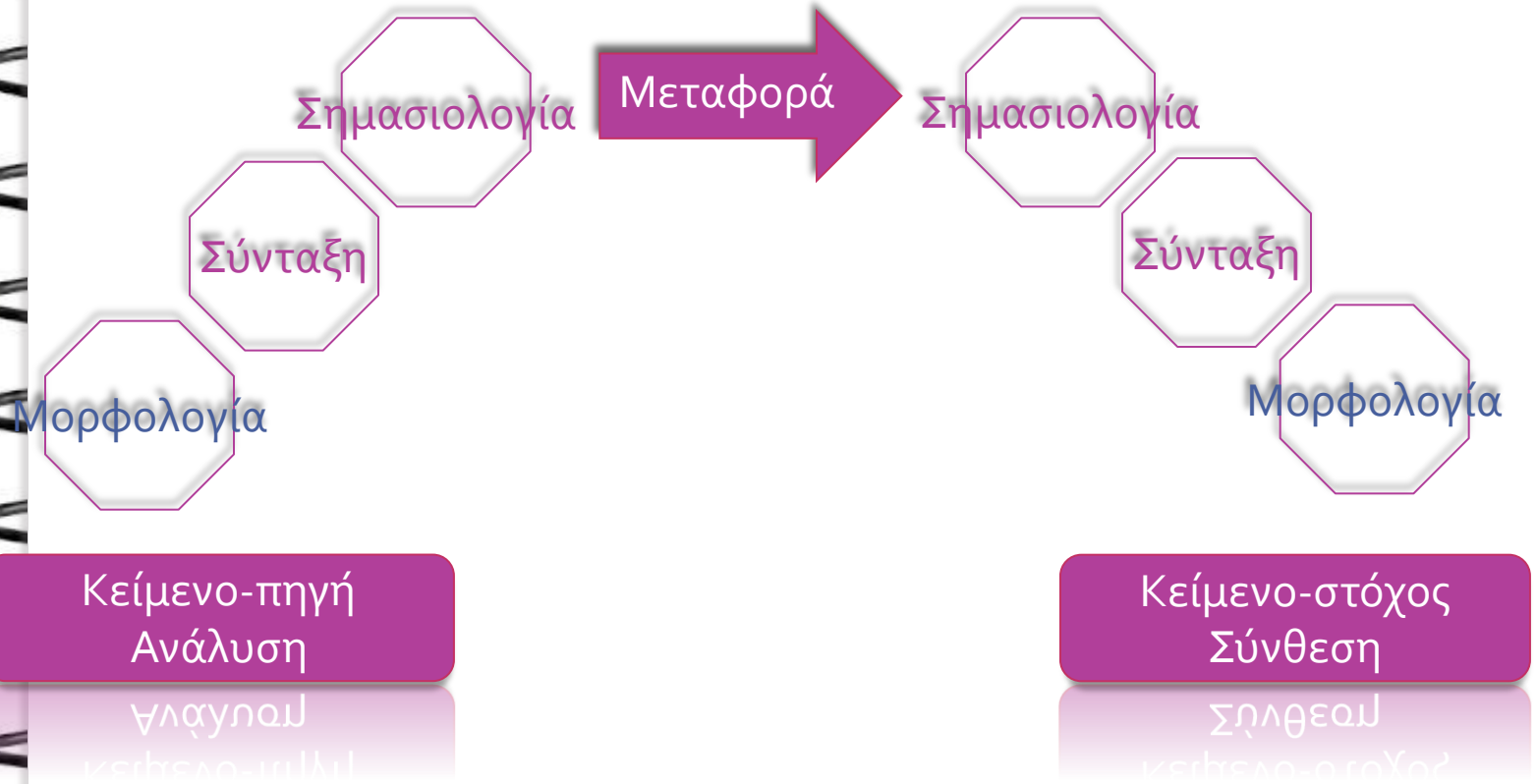
http://translate.google.com/

Μηχανική μετάφραση (ΜΜ)

- Με τον όρο **αυτόματη ή μηχανική μετάφραση** (αγγλ. machine translation) (εφεξής ΜΜ). **ορίζεται η αυτοματοποιημένη διαδικασία κατά την οποία μεταφέρεται ο γραπτός λόγος από μια γλώσσα-πηγή σε μια γλώσσα-στόχο**
- Η μετάφραση παράγεται **αυτόματα** από ένα **υπολογιστικό σύστημα** το οποίο διαχειρίζεται εξολοκλήρου τη μεταφραστική διαδικασία, χωρίς ανθρώπινη παρέμβαση (Σταύρου και Τζεβελέκου 2000, Σοφιανόπουλος 2009).
- Στη ΜΜ ο υπολογιστής δεν προσφέρει μόνο **μεταφράσεις λεξικών μονάδων**, αλλά παράγει **προτάσεις** ή **κείμενα**.
- Ο άνθρωπος χρήστης εισάγει ένα κείμενο πηγή στο σύστημα, πατάει το κουμπί για να ενεργοποιήσει τη λειτουργία της μετάφρασης και λαμβάνει στην έξοδο ένα κείμενο στόχο, χωρίς καθόλου να παρέμβει κατά τη διάρκεια της διαδικασίας
- Εδώ και δεκαετίες, η ΜΜ αποτελεί ένα από τα **δυσκολότερα επιστημονικά προβλήματα** της **υπολογιστικής γλωσσολογίας**, συγκεντρώνοντας το ενδιαφέρον επιστημόνων από διάφορους χώρους, όπως της πληροφορικής, της τεχνητής νοημοσύνης, της θεωρητικής και τυπικής γλωσσολογίας, της ψυχογλωσσολογίας, της λογικής και της φιλοσοφίας της γλώσσας.

Αυτόματη Μετάφραση

- Ruled-based machine translation (RBMT) – SYSTRAN



Αυτόματη Μετάφραση

- Statistical Machine Translation (SMT)

Προβλήματα:

- Ευθυγράμμιση προτάσεων (Sentence alignement)
- Σύνθετες λέξεις και ιδιωματικές εκφράσεις
- Μορφολογικοί τύποι
- Σύνταξη

Μηχανική Μετάφραση (Machine Translation)

- Από τις πρώτες εφαρμογές της υπολογιστικής γλωσσολογίας (1950)
- Τεράστιες εμπορικές εφαρμογές
  - Η ΕΕ ξοδεύει πάνω από 1 δις € σε κόστη μετάφρασης κάθε χρόνο
- Πολύ δύσκολο πρόβλημα ειδικά για μετάφραση:
  - εντελώς αυτοματοποιημένη
  - πραγματικού χρόνου
  - ανοιχτού λεξιλογίου

Εμπορικό αλλά και ερευνητικό ενδιαφέρον

- Συνδυάζει πολλές τεχνολογίες επεξεργασίας φυσικής γλώσσας:
  - Αναγνώριση μερών του λόγου
  - Συντακτική ανάλυση
  - Σύνθεση
  - Άρση αμφισημίας λέξης
  - Αναγνώριση ονομάτων –οντοτήτων
  - Επίλυση ασάφειας αναφορών
  - Κατανόηση φυσικής γλώσσας
  - Αναπαράσταση γνώσης του κόσμου

Χρησιμότητα Μηχανικής Μετάφρασης

- Σε εργασίες όπου μία πρόχειρη μετάφραση είναι επαρκής
  - Μετάφραση ιστοσελίδων
  - Διαγλωσσική ανάκτηση πληροφορίας
- Σε εργασίες όπου μπορεί να γίνει διόρθωση της αυτόματης μετάφρασης από κάποιον άνθρωπο-ειδικό
  - Human-assisted machine translation
- Σε εργασίες όπου επεξεργάζονται υπογλώσσες
  - Δελτία καιρού
  - Εγχειρίδια συσκευών

Προκλήσεις στην Αυτόματη Μετάφραση

Οι φυσικές γλώσσες διαφέρουν σε πολλά μεταξύ τους

- Μορφολογικές διαφορές
  - the → ο, η, το, τα, του, της, των, ...
- Αντωνυμίες
  - Σε πολλές γλώσσες (μορφολογικά πλούσιες) η αντωνυμία-υποκείμενο στην πρόταση εννοείται και τα μορφολογικά της χαρακτηριστικά καθορίζονται από την μορφολογία του ρήματος
  - Η κατάληξη του ρήματος στα Ισπανικά δείχνει ποιά αντωνυμία εννοείται
    - -o = I
    - -as = you
    - -a = he/she/it !!! (Ποιο θα επιλεγεί;)
    - -amos = we
    - -an = they

Προκλήσεις στην αυτόματη Μετάφραση

- Συντακτικές διαφορές (διάταξη των όρων)
  - language use → χρήση γλώσσας
    english(N1 N2) → greek(N2 N1)
  - the new house → la casa nueva
    english(DT J N) → spanish(DT N J)
  - IBM bought Lotus → IBM Lotus bought
    english(SUBJ V OBJ) → japanese(SUBJ OBJ V)
- Διαφορές στην έκφραση
  - Αγγλικά: *I am hungry* (είμαι πεινασμένος)
  - Γερμανικά: *Ich habe Hunger* (έχω πείνα)
  - Ελληνικά: *Πεινάω*

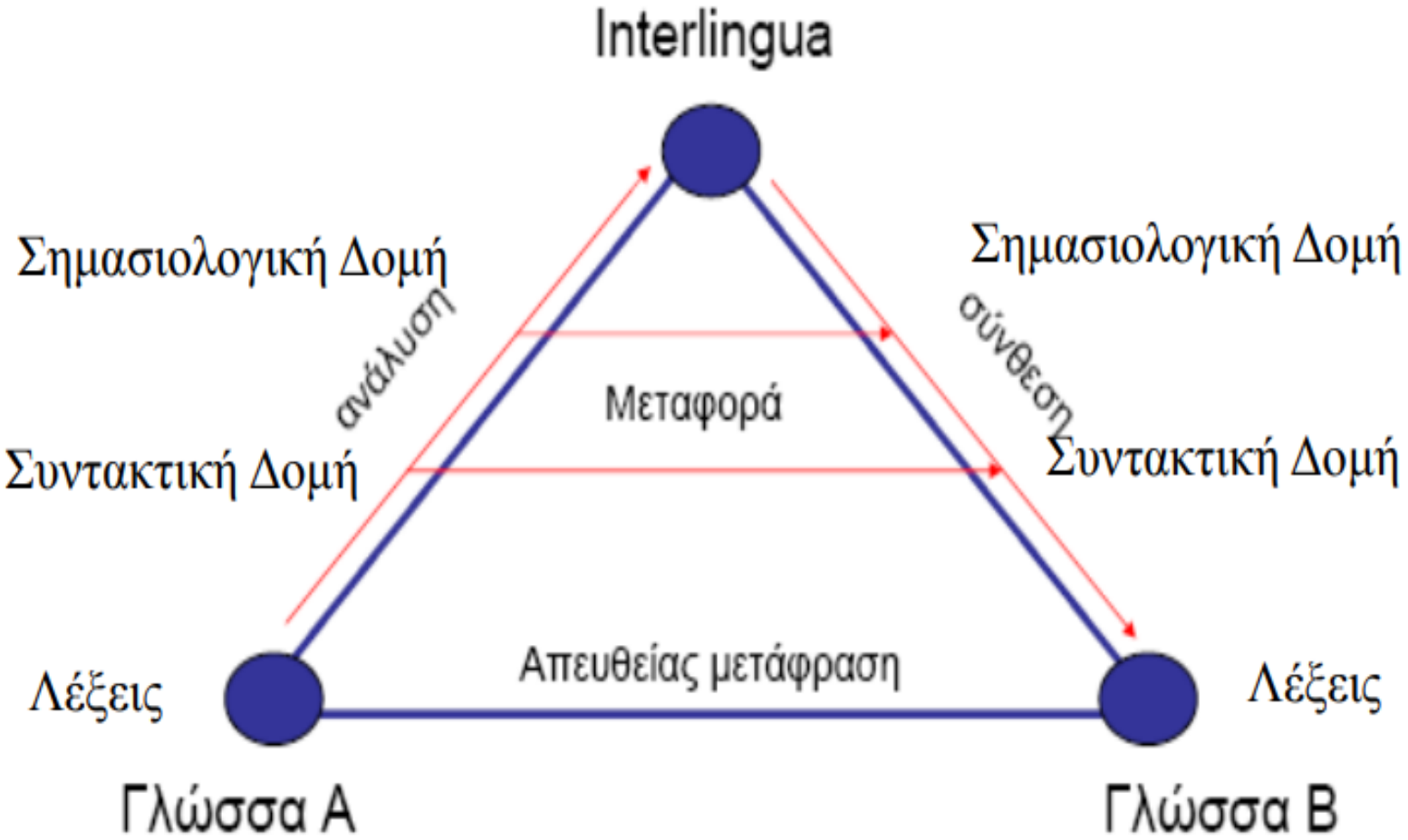
Προκλήσεις στην Αυτόματη Μετάφραση

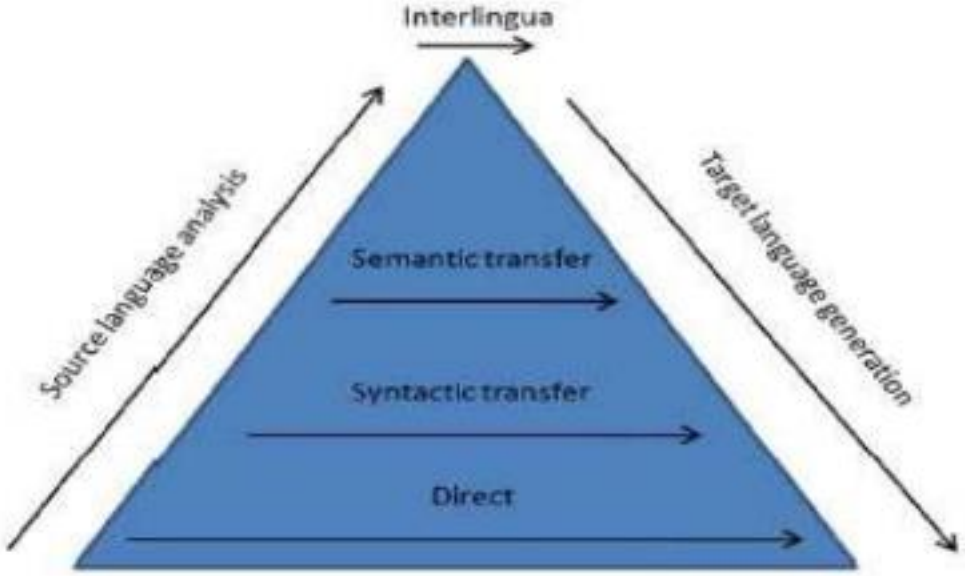
- Ο χρόνος των ρημάτων
  - I have been playing the piano for three years
  - Παίζω πιάνο τρία χρόνια
- Ιδιωματισμοί
  - He kicked the bucket → Πέθανε
  - She has always been a lame duck → Πάντα ήταν άχρηστη/βαρετή/ανίκανη

Κλασσικά Μοντέλα Μετάφρασης

- Interlingua
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β χρησιμοποιείται ως ενδιάμεσο μία ουδέτερη γλώσσα (interlingua - αναπαράσταση νοήματος)
- Transfer (μεταφορά)
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β ορίζεται μία διαδικασία ανάλυσης, μεταφοράς και σύνθεσης
- Direct (word-for-word) translation (απευθείας μετάφραση)
  - Για τη μετάφραση από μία γλώσσα Α σε μία γλώσσα Β γίνεται απευθείας μεταφορά από την μία στην άλλη

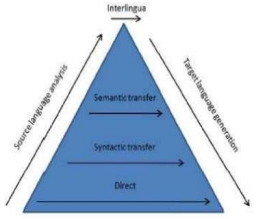
1. Τρίγωνο Vauquois





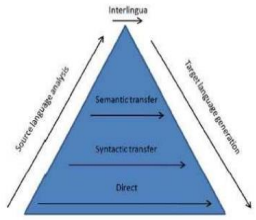
Η αριστερή πλευρά του τριγώνου αντιστοιχεί στη διαδικασία ανάλυσης του κειμένου στη γλώσσα-πηγή **ξεκινώντας από τη βάση του τριγώνου** και καταλήγοντας στην **κορυφή**, ενώ η δεξιά πλευρά αντιστοιχεί στη διαδικασία παραγωγής του κειμένου στη γλώσσα-στόχο, η οποία ξεκινά από την κορυφή της πυραμίδας και καταλήγει στο κάτω δεξιό άκρο.

Η πυραμίδα (τρίγωνο) τοι Vauquois



- Όσον αφορά την ανάλυση της γλώσσας, σύμφωνα με το τρίγωνο του Vauquois υπάρχουν **τρία διαφορετικά επίπεδα ανάλυσης**: η μορφολογία, η σύνταξη και η σημασιολογία.
- Ξεκινώντας από το κάτω μέρος του τριγώνου, **το πρώτο επίπεδο που συναντάμε είναι το μορφολογικό**, όπου η ανάλυση περιορίζεται στο επίπεδο της λέξης.
- Αμέσως παραπάνω, βρίσκεται η **ανάλυση σε επίπεδο δομών**, η οποία λαμβάνει υπόψη πληροφορίες που ξεπερνούν τα όρια της λέξης.
- Το τρίτο επίπεδο αντιστοιχεί στη σημασιολογική ανάλυση, η οποία προκύπτει μετά την ολοκλήρωση της αναγνώρισης της συντακτικής θέσης των συστατικών των προτάσεων του κειμένου στη γλώσσα-πηγή.

Η πυραμίδα (τρίγωνο) του Vauquois



- Η αντιστοίχιση των γλωσσικών δεδομένων από τη μια γλώσσα στην άλλη μπορεί να πραγματοποιηθεί σε κάθε επίπεδο. Αν η μετάφραση γίνει λέξη προς λέξη (αγγλ. direct/literal translation) χωρίς να προηγηθεί συντακτική ανάλυση του κειμένου στη γλώσσα-πηγή, τότε η αντιστοίχιση γίνεται στο κατώτατο επίπεδο της γλώσσας (βέλος Direct).
- Αν για τη μετάφραση λαμβάνονται υπόψη συντακτικοί κανόνες των δύο υπό μελέτη γλωσσών, τότε η αντιστοίχιση γίνεται μετά την ολοκλήρωση της συντακτικής ανάλυσης του κειμένου στη γλώσσα-πηγή (βέλος Syntactic transfer). Σε αυτή την περίπτωση, έχει ήδη ολοκληρωθεί η μορφολογική ανάλυση.
- Ομοίως, το βέλος Semantic transfer αντιστοιχεί στη μετάφραση που έχει γίνει, ενώ έχει ήδη προηγηθεί σημασιολογική ανάλυση του κειμένου στη γλώσσα-πηγή.
- Στην **κορυφή του τριγώνου** υπάρχει ο όρος **interlingua** ο οποίος αναφέρεται σε ένα θεωρητικό κωδικοποιημένο γλωσσικό σύστημα που είναι κοινό για όλες τις φυσικές γλώσσες και το οποίο χρησιμεύει ως **ενδιάμεση τεχνητή γλώσσα** για τη μεταφορά του μηνύματος από μια φυσική γλώσσα σε μια άλλη.

Στατιστική Μηχανική Μετάφραση

On voit Jon à la télévision

	good English? P(E)	good match to French? P(F|E)
Jon appeared in TV.		✓
Appeared on Jon TV.		✓
In Jon appeared TV.		
Jon is happy today.	✓	
Jon appeared on TV.	✓	✓
TV appeared on Jon.	✓	
TV in Jon appeared.		
Jon was not happy.	✓	

Ευθυγράμμιση - Alignment

- Παράλληλα κείμενα
  - Τα ίδια κείμενα γραμμένα στις δύο γλώσσες
- Επιπλέον, τα κείμενα πρέπει να είναι ευθυγραμμισμένα (aligned)
  - Σε ποια πρόταση (ή προτάσεις) μιας γλώσσας αντιστοιχεί μια πρόταση της άλλης γλώσσας
  - Σε ποια λέξη/φράση μιας γλώσσας αντιστοιχεί μια λέξη/φράση της άλλης γλώσσας

Στοίχιση κειμένων (alignment)

- Όσον αφορά τη στοίχιση κειμένων, οι δυσκολίες που αντιμετωπίζει το σύστημα είναι δύο:

α) η κατάτμηση των δύο κειμένων και

β) η στοίχιση των τμημάτων κειμένου.

Παραδείγματα Παράλληλων Κειμένων

- Πρακτικά Καναδικής Βουλής
- Επίσημη Εφημερίδα Ευρωπαϊκής Ένωσης
- Αναφορές Ηνωμένων Εθνών
- Εγχειρίδια χρήσης συσκευών
- Νομοθεσία Hong-Kong, Macao
- ...

Αξιολόγηση ΜΜ

- Τα συστήματα ΜΜ μπορούν να έχουν καλύτερα αποτελέσματα όταν μεταφράζουν **σύντομα** και **τυποποιημένα κείμενα**.
- Επιπλέον, η ΜΜ είναι αρκετά πιο αποτελεσματική στην **τεχνική** και **επιστημονική μετάφραση**, παρά στην οικονομική, νομική ή λογοτεχνική μετάφραση.
- Αντίθετα, τα συστήματα ΜΜ **αδυνατούν** να αποδώσουν **ικανοποιητικά αποτελέσματα** σε κείμενα με δημιουργικό, αισθητικό ή καλλιτεχνικό χαρακτήρα (λογοτεχνικά, ποιητικά, διαφημιστικά, χιουμοριστικά), καθώς δεν μπορούν να αποδώσουν τις λεπτές νοηματικές αποχρώσεις ούτε τις ιδιαιτερότητες στο ύφος (π.χ. ειρωνεία, χιούμορ) ή στο επίπεδο λόγου.

Κριτική και Συμπεράσματα

- Είναι αξιοσημείωτο ότι ορισμένα από τα βασικότερα προβλήματα της ΜΜ παραμένουν άλυτα:

1. επίλυση αμφισημιών,
2. λανθασμένη επιλογή λέξεων γλώσσας-στόχου,
3. επιλογή γένους αντωνυμιών και άρθρων,
4. διατήρηση συντακτικών σχημάτων γλώσσας - πηγής,
5. προβλήματα συμφωνίας όρων της πρότασης,
6. προβλήματα με προτάσεις που περιέχουν δευτερεύουσες προτάσεις κ.ο.κ.

Κριτική και Συμπεράσματα

- Είναι πράγματι **περίεργο** πώς μετά από 50 χρόνια έρευνας στον χώρο της ΜΜ υπάρχουν εμπορικά συστήματα τα οποία εξακολουθούν να παράγουν λανθασμένη μορφολογία, λανθασμένη συμφωνία όρων πρότασης ή να τοποθετούν τα ρήματα στην αρχή ή στο τέλος της πρότασης

Ευχαριστώ για την προσοχή σας!