Γλώσσες Προγραμματισμού ΙΙ

ΚΕΦΑΛΑΙΟ 1
(Εισαγωγή)

Επικοινωνία

Η **επικοινωνία** αντιστοιχεί σε τόσο ευρύ γνωστικό πεδίο και χρησιμοποιείται ως όρος σε πολλές περιοχές του επιστητού που γίνεται δύσκολος ένας ακριβής ορισμός της. Με κάποια γενική προσέγγιση θα λέγαμε ότι *"επικοινωνία είναι η ηθελημένη ανταλλαγή πληροφορίας και πραγματώνεται από παραγόμενα και κατανοούμενα σημεία (signs)επιλεγόμενα από σύστημα συμβατικών σημείων"*.

Οι άνθρωποι, όπως άλλωστε όλα τα ζώα, χρησιμοποιούν ως μέσο επικοινωνίας περιορισμένο σύνολο σημείων (χαμόγελο, χειραψία, χειρονομίες, κινήσεις σώματος κλπ). Επιπλέον ανέπτυξαν ένα πολύπλοκο και δομημένο σύστημα σημείων, τη **γλώσσα**, για να κοινωνήσουν ό,τι γνωρίζουν για τον κόσμο.

> *Τα δελφίνια και οι χιμπατζήδες χρησιμοποιούν λεξιλόγιο αποτελούμενο από δεκάδες σημεία και έχουν την ικανότητα να τα συμπαραθέτουν για να 'εκφραστούν'. Ο άνθρωπος είναι το μόνο είδος που μπορεί αξιόπιστα να επικοινωνήσει χρησιμοποιώντας απέραντο πλήθος μηνυμάτων, ποιοτικά διαφορετικών μεταξύ τους. Ενώ, για παράδειγμα, η ουρά της μέλισσας με το σκίρτημά της προσδιορίζει τη γωνία από τον ήλιο και την απόσταση του σημείου όπου βρίσκεται η τροφή, δημιουργώντας θεωρητικά άπειρο πλήθος μηνυμάτων, αλλά όχι ποιοτικά διαφορετικών.*

Γλώσσα

Οι *λέξεις*, που αποτελούν θεμελιώδεις νοηματικές μονάδες της γλώσσας, συνδέονται με *σημασίες* ('meanings') τις οποίες κοινωνούν οι άνθρωποι που επικοινωνούν μ' αυτές. Αυτό επιτελείται κατά σύμβαση δηλούμενη ή υπονοούμενη.

Το πιο απλό παράδειγμα αυτών των συμβάσεων αποτελούν τα *κύρια ονόματα* που αντιστοιχούν συνήθως σε άτομα (πρόσωπα, πόλεις, τοπωνύμια κλπ). Άλλο παράδειγμα αποτελούν τα *ονόματα* γενικά που αντιστοιχούν στα διάφορα είδη των αντικειμένων και τα *ρήματα* που αντιστοιχούν σε δράσεις (π..χ. 'δίνω το ρολόι του Γιώργου'), σε καταστάσεις (π.χ. 'η γάτα κοιμάται') ή σε σχέσεις μεταξύ των αντικειμένων (π.χ. 'αιτία του Τρωικού πολέμου στάθηκε η ωραία Ελένη').

Η αντιστοιχία αυτή δεν μοναδική (ένα προς ένα) ούτε αναγκαία. Αυτό αποδεικνύεται με τον πλούτο των γλωσσών του ανθρώπου.

Τέλος, η 'κατανόηση' της γλώσσας είτε από τον άνθρωπο είτε από τη μηχανή στηρίζεται στην κοινωνούμενη αυτή σύμβαση, που προαναφέρθηκε. Επομένως τέτοιες κοινωνούμενες συμβάσεις είναι απαραίτητο να δοθούν στις μηχανές (τους υπολογιστές) ώστε να επιτελέσουν την επεξεργασία της φυσικής γλώσσας.

Υπολογιστική Γλωσσολογία ή Επεξεργασία Φυσικής Γλώσσας

Η επιστημονική περιοχή που φέρει αυτόν τον τίτλο σε αδρές γραμμές καλύπτει τις ακόλουθες κατευθύνσεις.

- Σχεδιασμός και υλοποίηση υπολογιστικών **μοντέλων** της φυσικής γλώσσας, και πιο συγκεκριμένα:
  την *αναγνώριση* ή *κατανόηση* της φυσικής γλώσσας από υπολογιστή, και
  την *παραγωγή* φυσικής γλώσσας ή *σύνθεση*.

- Οι εφαρμογές της περιοχής αυτής αγκαλιάζουν:
  τον *διάλογο* με τον υπολογιστή
  την *μηχανική μετάφραση* (από γλώσσα σε γλώσσα, από γλώσσα σε βάση δεδομένων και αντιστρόφως κλπ), και
  την *αναδίφηση* (browsing) ή την *διήθηση* (filtering) κειμένων φυσικής γλώσσας από έναν πράκτορα ή παράγοντα (agent).

Ο όρος **Επεξεργασία Φυσικής Γλώσσας** (Natural Language Processing) είναι παραδοσιακός και προέρχεται από τον ομώνυμο κλάδο της *Τεχνητής Νόησης (Artificial Intelligence)*. Σήμερα αντιστοιχεί στην *εφαρμοσμένη Υπολογιστική Γλωσσολογία*.

Υπολογιστική Γλωσσολογία

Η μελέτη της γλώσσας αποτελεί αντικείμενο της κλασικής **Γλωσσολογίας**. Η **Υπολογιστική Γλωσσολογία** ως διεπιστημονική γνωστική περιοχή βρίσκεται ανάμεσα στην ***Γλωσσολογία*** και την ***Πληροφορική***. Ανήκει στις *Γνωστικές (Cognitive) Επιστήμες* και επικαλύπτεται με την *Τεχνητή Νόηση (Artficial Intelligence)*, κλάδο της Πληροφορικής (Computer Science), που ασχολείται με τα υπολογιστικά μοντέλα της ανθρώπινης γνώσης.

Η Υπολογιστική Γλωσσολογία έχει *θεωρητικό* και *πρακτικό* κλάδο. Η ***θεωρητική Υπολογιστική Γλωσσολογία*** πραγματεύεται τις τυπικές θεωρίες παράστασης της γλωσσικής γνώσης και τους φορμαλισμούς που έχουν ανάγκη οι άνθρωποι για να παράγουν και να κατανοήσουν τη γλώσσα. Οι θεωρίες αυτές σήμερα έχουν φτάσει σε τέτοια πολυπλοκότητα που μόνο με υπολογιστές μπορεί κανείς να τις διαχειριστεί. Οι επιστήμονες του κλάδου αυτού αναπτύσσουν *τυπικά μοντέλα* που εξομοιώνουν απόψεις της ανθρώπινης γλωσσολογίας και τα υλοποιούν σε υπολογιστικά συστήματα. Αυτά αποτελούν τη βάση αποτίμησης και περεταίρω ανάπτυξης των αντίστοιχων θεωριών. Στην περιοχή αυτή της επιστήμης σημαντικό ρόλο παίζουν ευρήματα από την *Γνωστική Ψυχολογία*. Επίσης, η *Ψυχογλωσσολογία*, κλάδος της *Ψυχολογίας*, εξετάζει τις γνωστικές διαδικασίες που συνιστούν την χρήση της ανθρώπινης γλώσσας.

Η ***εφαρμοσμένη Υπολογιστική Γλωσσολογία*** εστιάζεται στα πρακτικά αποτελέσματα της μοντελοποίησης της χρήσης της ανθρώπινης γλώσσας. Οι μέθοδοι, οι τεχνικές, τα εργαλεία και οι εφαρμογές αυτής της περιοχής συχνά συμπεριλαμβάνονται στον όρο (ανθρώπινη) **γλωσσική μηχανική** (language engineering) ή **γλωσσική τεχνολογία** (language technology). Παρότι τα σημερινά υπολογιστικά συστήματα απέχουν πολύ από τις ανθρώπινες ικανότητες έχουν πολλές πιθανές εφαρμογές. Στόχος είναι η ανάπτυξη συστημάτων με κάποια γνώση της ανθρώπινης γλώσσας. Τα συστήματα αυτά χρησιμοποιούνται στην καλυτέρευση της επικοινωνίας ανθρώπου - μηχανής. Οι σημερινοί υπολογιστές δεν ‘καταλαβαίνουν’ τη γλώσσα μας αλλά, από την άλλη πλευρά, οι (τεχνητές) *γλώσσες προγραμματισμού* είναι δύσκολο να μαθευτούν και δεν αντιστοιχούν στη δομή της ανθρώπινης σκέψης.

Η παράδοση της Επεξεργασίας Φυσικής Γλώσσας

Με τους πρώτους υπολογιστές (1950 - 1960) ο χειρισμός **αριθμών** και στη συνέχεια **λέξεων** έδωσε τη δυνατότητα μιας πρώτης 'ηλεκτρονικής' παράστασης και επεξεργασίας γλωσσικής γνώσης. Αργότερα, η ***μετρική*** και η **στατιστική** του κειμένου επέτρεψαν την καλλιέργεια της **υφολογίας** (style analysis) και της **εύρεσης** της **ταυτότητας** του **συγγραφέα** ή συντάκτη ενός κειμένου. Ακολούθησε η αυτόματη δημιουργία **δεικτών** (indices) και η εύρεση των **λέξεων** μέσα **στα συμφραζόμενά** τους (keywords in context: concordance lists). Σήμερα η περιοχή αυτή εξακολουθεί να καλλιεργείται και ονομάζεται *Literary and Linguistic Computing*. *(Την αποδίδουμε ελεύθερα ως Υπολογιστική Γραμματεία και Γλωσσολογία).* Η **μηχανική** (και όχι αυτόματη) **μετάφραση** (machine translation) έδωσε τα πρώτα δείγματα στα μέσα της δεκαετίας του '60.

Ο T. Winograd το 1972, με το σύστημα *SHRDLU*, γραμμένο σε Lisp, προσέφερε το πρώτο **σύστημα κατανόησης φυσικής γλώσσας** (natural language understanding) θεμελιώνοντας ουσιαστικά την "περί υπάρξεως απόδειξη" της περιοχής. Ερμηνεία ερωτήσεων, δηλώσεων και οδηγιών, ικανότητα συνεπαγωγής, εκμάθηση νέων λέξεων αποτέλεσαν μερικές από τις αρετές του συστήματος.

Η **διαδικασιακή** (procedural) ήταν η πρώτη μορφή *παράστασης γνώσης*. Ακολούθησε η **δηλωτική** (declarative). Η διαδικασιακή επιχειρεί να περιγράψει την κίνηση ή τη δράση και φιλοδοξεί να παραστήσει τη γνώση (γλωσσική ή μη) με *διαδικασίες*, με άλλα λόγια το "**πώς**" συμβαίνει η γνώση. Η δηλωτική επιχειρεί την παράσταση των *συμβολικών δομών* καθαυτών, μ' άλλα λόγια προσπαθεί να περιγράψει "**τι**" συμβαίνει. Ο δηλωτικός προγραμματισμός εξελίχθηκε και συμπεριέλαβε **κανόνες συνεπαγωγής** (inference rules). Ο προγραμματισμός αυτός λέγεται **Λογικός Προγραμματισμός** και η σημαντικότερη αντίστοιχη γλώσσα είναι η **Prolog**.

Η επιβολή της Δομής

Τα **αυτόματα πεπερασμένων καταστάσεων** (finite state automata: FSA) επιστρατεύτηκαν για να αναγνωριστούν γλωσσικές οντότητες (π.χ. λεξικά στοιχεία, μέρη του λόγου) και τα **δίκτυα μεταφοράς πεπερασμένων καταστάσεων** (Finite State Transition Networks: FSTN) για να δημιουργήσουν γλωσσικές οντότητες. Σημαντικό σημείο αναφοράς στην περιοχή αυτή αποτελεί το πρόγραμμα *ELIZA* του J. Weizenbaum που βασίστηκε σ΄αυτά, το 1960.

Τα **αναδρομικά δίκτυα μεταφοράς** (Recursive Transition Networks: RTN) που ακολούθησαν ήσαν δίκτυα μεταφοράς πεπερασμένων καταστάσεων (FSTN) με *αναδρομή*. Μπορούσαν να παραστήσουν γραμματικές του τύπου $a^n b^n$ και τις παραστάσεις με παρενθέσεις.

Τα **διευρυμένα δίκτυα μεταφοράς** (Augmented Transition Networks: ATN) είχαν επιπλέον και *μνήμη* για να μην επαναλαμβάνουν διαδρομές. Παριστάνοντας δε μια μορφή *δράσης* ή *ενέργειας* (action) κατέστρεφαν τον δηλωτικό χαρακτήρα τους.

Οι πρώτοι "**τεχνολογητές χάρτη**" ή "**συντακτικοί αναλυτές χάρτη**" (chart parsers) που κατέγραφαν τις πετυχημένες απόπειρες, για να μην επαναλαμβάνουν άσκοπα τους ελέγχους νομιμότητας, επιβιώνουν μέχρι σήμερα ως ιδέα με διαφορετικές μεθόδους υλοποίησης.

Με τη γέννηση της Prolog και της *λογικής συνεπαγωγής (inference)* που αυτή επιτρέπει δημιουργήθηκαν οι πρώτοι Τεχνολογητές (parsers) **γραμματικών ορισμένης πρότασης** (Definite Clause Grammars: DCG). Αυτοί είναι Τεχνολογητές **γραμματικών φραστικής δομής** (Phrase Structure Grammars: PSG) σε γλώσσα Prolog, και αναμιγνύουν διαδικασιακή με δηλωτική μορφή παράστασης.

Η **αμφισημία** ως εγγενές στοιχείο της φυσικής γλώσσας που απαντάται σε όλα τα επίπεδα μελέτης της φυσικής γλώσσας, αποτελεί 'πονοκέφαλο' της μηχανικής επεξεργασίας φυσικής

γλώσσας. Έρχεται σε αντίθεση με τον *ντετερμινισμό* των υπολογιστικών μηχανών και τις ντετερμινιστικές φιλοσοφικές δοξασίες. Η άρση της αμφισημίας αποτέλεσε και αποτελεί πρόκληση στην έρευνα της Γλωσσολογίας και της Υπολογιστικής Γλωσσολογίας.

Η παράσταση της Σημασίας

Η **κατανόηση της γλώσσας** (language understanding) με μια λέξη αποτελεί την αντιστοίχιση *σημασίας* (meaning) στις *γλωσσικές μορφές*, ενώ το αντίθετο ονομάζεται **γένεση γλώσσας** (language generation).

Ποια όμως μπορεί να είναι η *δομή* που 'παριστάνει' τη *σημασία* (meaning); Η Τεχνητή Νόηση (Artificial Intelligence: AI) θεωρώντας τον νου ως υπολογιστική διάταξη πρότεινε την *διαδικασία* (procedure). Οι λογικοί (logicians) προτείνουν τη σημασία απλών προτάσεων.

Γεννιέται έτσι η λεγόμενη **διαδικασιακή σημασιολογία** (procedural semantics). Η σημασία μιας εντολής ή οδηγίας (command) είναι αυτή καθαυτή η διαδικασία που ακολουθείται για να υλοποιήσει την απαιτούμενη ενέργεια / δράση (action). Παρόμοια με την εντολή που αντιστοιχίζεται έτσι στη δράση, η ερώτηση αντιστοιχίζεται στην απάντηση και η δήλωση (statement) στην νέα πληροφορία (new information) που αυτή εμπεριέχει.

Υπάρχουν όμως όρια στην παράσταση της γνώσης μόνο με τις διαδικασίες. Η διαδικασία μπορεί μόνο να "τρέξει" (run), μπορεί μόνο να εκτελεστεί.

Με τα **σημασιολογικά δίκτυα** (semantic networks) ο Schank το 1960 βασίζεται στη θεωρία των Katz και Fodor (1963) και εισηγείται 11 *πρωταρχικά στοιχεία* ή *αρχέτυπα* (primitives) τα οποία ισχυρίζεται ότι μπορούν να συνδυαστούν για να εκφράσουν όποιο γεγεονός του κόσμου. Η άποψή του για τις **εννοιακές εξαρτήσεις** (conceptual dedpendencies) παριστάνεται με *γραφήματα* (graphs) όπου *κόμβοι* (nodes) είναι οι έννοιες και οι μεταξύ τους *σύνδεσμοι* (links) οι σχέσεις εξάρτησης (dependencies).

Ακολουθούν το 1975 τα **πλαίσια** (frames) του Minsky. Αυτά προσφέρουν την διαδικασία που ενεργοποιείται από δεδομένα (data driven processing) που καταχωρούνται σε ειδικές θέσεις (slots) μέσα σ' αυτά.

Το 1970 το πρόγραμμα *LUNAR* του Woods, που απαντά σε ερωτήσεις για σεληνιακά πετρώματα, αποτελεί ένα από τα πρώτα συστήματα που χρησιμοποιούν τη *λογική πρώτης τάξης* (first order logic: FOL) για την παράσταση γνώσης.

Η *τυπική σημασιολογία* (formal semantics) μαζί με τη *φιλοσοφία της γλώσσας* μας χαρίζουν την ιδέα πως η σημασία είναι **συνθετική** (compositional semantics) με απαρχή τις ιδέες του Frege (1890!). Σύμφωνα μ' αυτή την άποψη η σημασία μιας πρότασης συναπαρτίζεται από τις επιμέρους σημασίες των συστατικών λέξεών της. Ο Montague (1974) εργάστηκε στην αρχή της συνθετικότητας για να αναλύσει προτάσεις φυσικής γλώσσας με αρκετή επιτυχία.

Πολλά υπολογιστικά συστήματα αξιοποίησαν τις ιδέες των Katz και Fodor προκειμένου να άρουν τις αμφισημίες με τη μέθοδο των **επιλεγομένων περιορισμών** (selectional restrictions). Σύμφωνα μ' αυτή κάθε σημασία μιας λέξης έχει και ένα *σημάδι* (marker). Η σημασία της πρότασης προκύπτει από τη συμφωνία μεταξύ των σημαδιών των συστατικών της λέξεων.

Ο ρόλος της Γνώσης

Η **παράσταση της γνώσης** (knowledge representation) ήταν και είναι ακόμη μια από τις πιο ενδιαφέρουσες απόπειρες άρσης της αμφισημίας. Το σύστημα κατανόησης μιας πρότασης φυσικής γλώσσας θα πρέπει να διαθέτει γνώση για τον (περιορισμένο) κόσμο όπου αυτή η πρόταση βρίσκεται / εκφέρεται.

Οι **γραφές** (scripts) του Schank (1977) εξετάζουν την *αιτιότητα* και την *διαδοχή* των γεγονότων. Κατά την άποψη αυτή μπορούμε να παραστήσουμε τη γνώση μας για πρωτοτυπικά γεγονότα ως ακολουθίες **προσδοκομένων ενεργειών** / **δράσεων** (expected actions) σε μια γραφή (script) για τα γεγονότα. Σε κάθε πρωτότυπο στοιχείο γνώσης υποδεικνύεται κι ένα *μοντέλο αναδίφησης* (retrieval model). Θεμελιώνεται έτσι η κατανόηση που βασίζεται στις **προσδοκίες** (expectation based understanding). Πώς όμως να χειριστούμε αποκλίσεις από τις προσδοκόμενες ενέργειες;

Η νέα Γλωσσική Τεχνολογία

Σταχυολογούμε επιγραμματικά σ' αυτό το εισαγωγικό σημείωμα μερικές από τις προόδους της Επεξεργασίας της Φυσικής Γλώσσας σε διαφόρους τομείς εφαρμογής.

Ο "**προγραμματισμός**" (planning) εφαρμόστηκε στη ρομποτική σε συστήματα που αξιοποιούν στοιχεία κατανόησης φυσικής γλώσσας στην παραγωγή.

Η διαλογική συμπεριφορά του ανθρώπου μελετιέται διεπιστημονικά και αξιοποιούνται τ' αποτελέσματα της έρευνας στον σχεδιασμό **διεπικοινωνίας φυσικής γλώσσας** (natural language interface / front ends) με υπολογιστικά συστήματα και βάσεις δεδομένων.

Η **μηχανική μετάφραση** περιορισμένων και κυρίως τεχνικών και διοικητικών εκφράσεων είναι εφικτή σήμερα με τις νέες τεχνικές. Επίσης, η **άντληση γραμματικών στοιχείων** από *σώματα (corpora)*. Τα στοιχεία γλωσσικής γνώσης στους *επεξεργαστές κειμένου* αυξάνουν ποιοτικά. Τέλος, η **γένεση επεξήγησης** (explanation generation) σε συστήματα βάσεων γνώσης παρουσίασε πρόοδο (ειδικότερα σε συστήματα ιατρικής συμβουλευτικής).