ΕΛΕΤΟ – 7ο Συνέδριο «Ελληνική Γλώσσα και Ορολογία»
Αθήνα, 22-24 Οκτωβρίου 2009

Ατομική και συλλογική ευφυΐα στην ανάπτυξη γλωσσικών πόρων : Κανονιστικές και "κοινωνικές" προσεγγίσεις

Στέλιος Πιπερίδης

ΠΕΡΙΛΗΨΗ

Η παρούσα ομιλία διερευνά την ιστορική διαδρομή των μεθοδολογικών προσεγγίσεων ανάπτυξης γλωσσικών πόρων και γλωσσικής τεχνολογίας, συσχετίζοντας τις φάσεις εξέλιξης με τα ιστορικά αντίστοιχα ρεύματα στη φιλοσοφία της γλώσσας. Χρησιμοποιώντας ως παραδειγματικό τομέα τη μηχανική μετάφραση, παρουσιάζονται οι αλληλεπιδράσεις και συσχετίσεις από τα αρχικά στάδια μέχρι σήμερα. Στη συνέχεια παρουσιάζεται το τοπίο των γλωσσικών πόρων και της γλωσσικής τεχνολογίας σήμερα, μέσα από το πρίσμα τριών μεγάλης εμβέλειας δράσεων για τον επανακαθορισμό της έννοιας των γλωσσικών πόρων, των προτεραιοτήτων και του τρόπου οργάνωσης και χρήσης τους στο περιβάλλον του παγκόσμιου ιστού και των κοινωνικών δικτύων του.

Individual and collective intelligence in language resources development : Normative and social approaches

Stelios Piperidis

ABSTRACT

This talk investigates the different methodological approaches to language resources and technology development overtime, inter-relating their evolution with historically corresponding philosophical thought about language. Using machine translation as an exemplary application, the relative impact and interrelations from initial phases until today are presented. Next, we present the language resources and technology landscape today, from the viewpoint of three initiatives of major scope towards redefining the concept and extension of language resources, the priorities, the organizational models and ways of their use in the current World Wide Web environment and its social networks.

Συνοπτική περιγραφή

Η κωδικοποιημένη γλωσσολογική γνώση και τα γλωσσικά δεδομένα, σε διάφορες μορφές (σώματα κειμένων, λεξικά, ορολογικές βάσεις, οντολογίες, τυπικές γραμματικές, κλπ.) βρίσκονται στον πυρήνα της γλωσσικής τεχνολογίας τα τελευταία 60 χρόνια. Την περίοδο του παραδοσιακού κύκλου της τεχνητής νοημοσύνης, με τις τότε νομοθετικές (rule-based), γνωσιοκεντρικές (knowledge-based) και παραγωγικές (deductive) μεθόδους συμπερασματολογίας (inference) ως κυρίαρχη επιστημονική μέθοδο, οι τυπικές λεξιλογικές βάσεις δεδομένων, οι τυπικές γραμματικές και οι βάσεις πραγματολογικών δεδομένων συναποτελούσαν το βασικό κορμό γνώσης και δεδομένων. Αν ανατρέξει κανείς στις ιστορικές αναφορές εξέλιξης του πεδίου της μηχανικής μετάφρασης, θα διαπιστώσει την επικράτηση μεθόδων όπως της ευθείας λεξικής μετάφρασης (direct translation), των γραμματικών μεταφοράς (transfer grammars) και των μεθόδων μετάφρασης βάσει διαγλώσσας (interlingua). Ατομικά, αλλά και συνεργατικά, οι μέθοδοι αυτές προσπάθησαν να επιλύσουν προβλήματα αμφισημίας, και αργότερα γλωσσικής ποικιλίας, ψάχνοντας αποδεκτούς τρόπους μετάφρασης της, προβληματικής για τον Bar-Hillel, πρότασης «Little John's box is in the pen", μέχρι τα μέσα της δεκαετίας του '80.

Η τάση αυτή ήταν αναμενόμενη από επιστημολογική άποψη, αναπτυσσόμενη, υπό την άμεση επιρροή των επιλογών της αναλυτικής φιλοσοφίας της γλώσσας του πρώτου μισού του 20^{ού} αιώνα, κυρίως μέσω της αρχής της συνθετικότητας (compositionality) των εννοιών. Σε αυτή τη γραμμή σκέψης συνηγόρησαν ο Gottlob Frege με τον λογικισμό και αναγωγισμό του, ο Bertrand Russell με τον λογικό ατομισμό και τη λογική ανάλυση, και ο πρώιμος Wittgenstein, μέσα από το Tractatus, με την αντίληψη της γλώσσας ως καθρέφτη του κόσμου και το ρόλο της λογικής ως συνδετικού ιστού μεταξύ γλώσσας και πραγματικότητας.

Η γλωσσική και οντολογική σχετικότητα, η απροσδιοριστία της μετάφρασης, το πρόβλημα της αναφοράς, της ιδανικής-καθολικής γλώσσας, η αναπαραστατική λειτουργία της γλώσσας ήταν μερικά μόνο από τα προβλήματα που επηρέασαν τόσο την περαιτέρω εξέλιξη της συνθετικότητας και του αναγωγισμού, αλλά και την διασάφηση της διάσημης πλέον φράσης «box is in the pen».

Η σκέψη του ύστερου Wittgenstein, με τα γλωσσικά παιχνίδια (language games), την απόρριψη της αναπαραστατικότητας, και την εγκαθίδρυση του νοήματος ως χρήση, αλλά και η θέση περί απουσίας απόλυτου κριτηρίου συνωνυμίας του Willard Van Orman Quine, και η απόλυτη συσχέτιση νοήματος με τη χρήση των λέξεων και την αντίστοιχη συμπεριφορά των

φυσικών ομιλητών, οδήγησαν σε μια νέα στροφή τόσο τη φιλοσοφία γλώσσας, όσο και τη γλωσσική τεχνολογία και τα σχετικά με αυτήν πεδία, κυρίως της μηχανικής μετάφρασης.

Η ομιλία αυτή ασχολείται στο πρώτο μέρος της με την αλληλεπίδραση μεταξύ φιλοσοφικής σκέψης και επιστημονικής μεθόδου, εξετάζοντας τις σχέσεις, άμεσες ή έμμεσες, που μπορούν να οδηγήσουν σε καλύτερη κατανόηση και πρόβλεψη του κύκλου εξέλιξης της γλωσσικής τεχνολογίας και του θεωρητικού της υπόβαθρου.

Η σύγχρονη επικράτηση πραγματολογικών, μπεχαβιοριστικών προσεγγίσεων, της αντίληψης του νοήματος ως χρήση, αλλά και των συνεπαγόμενων επαγωγικών, τεχνολογικά στατιστικών, μεθόδων, συνεπικουρούμενη από την αλματώδη αύξηση του διαθέσιμου ψηφιακού περιεχομένου στον παγκόσμιο ιστό, διαμορφώνει μια καινούργια πραγματικότητα στο χώρο των γλωσσικών δεδομένων.

Μέσα στα διαθέσιμα 487 δις. GB ψηφιακού περιεχομένου που δημιουργήθηκε σε παγκόσμια κλίμακα το 2008, ιδιαίτερο ενδιαφέρον για τη σύγχρονη μεταφραστική τεχνολογία παρουσιάζουν τα παράλληλα κείμενα σε πλειάδα γλωσσικών ζευγών (με διαφοροποιούμενο όγκο ανά ζεύγος), παράλληλα κείμενα από τα οποία δημιουργούμε μεταφραστικά μοντέλα για συστήματα στατιστικής μηχανικής μετάφρασης και αντλούμε δίγλωσσα λεξιλογικά και ορολογικά δεδομένα για πλήθος θεματικών τομέων και κειμενικών ειδών. Μονογλωσσικά δεδομένα, σε οποιαδήποτε μορφή, από λίστες λέξεων μέχρι επικοινωνιακά δεδομένα διαλόγων ή άλλης μορφής ανταλλαγής πληροφορίας, χρησιμοποιούνται για την μοντελοποίηση του παραγόμενου μονογλωσσικού, μεταφρασμένου, ή μη, λόγου.

Στο δεύτερο μέρος της ομιλίας αυτής, θα συζητήσουμε το σύγχρονο τοπίο των γλωσσικών δεδομένων, των ιδιαιτεροτήτων που παρουσιάζουν στον σημερινό παγκόσμιο ιστό, και των εργαλείων που χρησιμοποιούνται, ή απαιτούνται για την επεξεργασία τους. Στο πλαίσιο αυτό θα παρουσιαστούν τρεις πρωτοβουλίες σε ευρωπαϊκό και παγκόσμιο επίπεδο: η ευρωπαϊκή ερευνητική υποδομή CLARIN (www.clarin.eu), το θεματικό δίκτυο FLARENET (www.flarenet.eu) και η σχεδιαζόμενη υποδομή ανοικτών γλωσσικών πόρων ORI – Open Resource Infrastructure.

Παρόλο που η αναγκαιότητα των γλωσσικών πόρων (γλωσσικών δεδομένων και εργαλείων) αναγνωρίζεται ευρύτατα, ταυτόχρονα εντοπίζεται ένα οξύ πρόβλημα: η αποσπασματικότητα στη δημιουργία τους και η έλλειψη ισχυρών διεθνών προτύπων για την περιγραφή ή την κωδικοποίησή τους οδηγούν στην αδυναμία επαναχρησιμοποίησης, διασύνδεσης και αξιοποίησης των γλωσσικών πόρων στην ανάπτυξη νέων τεχνολογιών και εφαρμογών.

Το έργο CLARIN (http://www.clarin.eu) είναι μία πανευρωπαϊκής εμβέλειας απόπειρα να συγκεντρωθούν, να συντονιστούν και τελικά να διατεθούν στην ερευνητική κοινότητα γλωσσικοί πόροι και υπηρεσίες μέσω μιας κατανεμημένης διαδικτυακής ερευνητικής υποδομής. Η πρωτοβουλία απευθύνεται κυρίως στην ερευνητική κοινότητα των ανθρωπιστικών και κοινωνικών επιστημών και στοχεύει να παράσχει πληροφορία σχετικά με την ύπαρξη γλωσσικών πόρων, συντονισμό της δημιουργίας, αρχειοθέτησης, διαχείρισης και πρόσβασης στους γλωσσικούς πόρους και παροχή υπηρεσιών μέσω γλωσσικών εργαλείων.

Το θεματικό δίκτυο FLARENET (http://www.flarenet.eu) αποτελεί ένα διεθνές φόρουμ, που απαρτίζεται από μια διαρκώς διευρυνόμενη κοινότητα, με σκοπό να διευκολύνει την επικοινωνία ανάμεσα στους οργανισμούς Γλωσσικών Πόρων και Γλωσσικής Τεχνολογίας, ώστε να δώσει νέα ώθηση και ταυτότητα στην ερευνητική κοινότητα, να διαμορφώσει ένα κοινό όραμα στον χώρο των γλωσσικών πόρων και της γλωσσικής τεχνολογίας για τα επόμενα χρόνια και να διευκολύνει την ανάπτυξη Ευρωπαϊκής στρατηγικής για την ενίσχυση του τομέα και της ανταγωνιστικότητας του σε Ευρωπαϊκό και παγκόσμιο επίπεδο. Το έργο υλοποιεί τους στόχους αυτούς εντοπίζοντας τομείς προτεραιοτήτων, προωθώντας διαδικασίες προτυποποίησης και υιοθέτησης βέλτιστων πρακτικών σε όλα τα επίπεδα ανάπτυξης ανάπτυξης πόρων και τεχνολογίας, και παρέχοντας συστάσεις για δράσεις τόσο στην Ευρωπαϊκή Επιτροπή όσο και σε εθνικές κυβερνήσεις και οργανισμούς.

Η σχεδιαζόμενη υποδομή ανοικτών γλωσσικών πόρων ORI – Open Resource Infrastructure, δράση στο πλαίσιο του σχεδιαζόμενου δικτύου αριστείας Technologies for Multilingual Europe, υιοθετεί μια εξελιγμένη άποψη για τους γλωσικούς πόρους, θεωρώντας τους δυναμικές, διαρκώς εξελισσόμενες, οντότητες. Οι γλωσσικοί πόροι δεν μπορούν πλέον να θεωρούνται μονολιθικά, στατικά αντικείμενα. Αποτελούν ζωντανές οντότητες, αναπόσπαστο τμήμα δεδομένων ανθρώπινης συμπεριφοράς, που μπορούν να τις επεξεργαστούν γλωσσικά και άλλα εργαλεία, και οι οποίες συσχετίζονται με άλλα αισθησιοκινητικά και συμπεριφοριστικά δεδομένα. Ως τέτοια, τα γλωσσικά δεδομένα, σχετικά δεδομένα σε άλλα μέσα (π.χ. ακίνητες ή κινούμενες εικόνες) και άλλες τροπικότητες (π.χ. χειρονομίες, νοήματα κλπ.), και εργαλεία και τεχνολογίες που επεξεργάζονται αυτά τα δεδομένα με σκοπό την επισημείωσή τους, την εξαγωγή γνώσης από αυτά, τον επαναπροσανατολισμό τους, τη δημιουργία νέων συνδέσεων και σχέσεων μεταξύ τους, την παραγωγή ή την αποκάλυψη νέας πληροφορίας και γνώσης, δημιουργούν ένα νέο ενιαίο σύνολο.

Και οι τρεις παραπάνω πρωτοβουλίες διατρέχονται κεντρικά από τις αρχές των κατανεμημένων υποδομών και των ανοικτών πόρων, ενώ αντιμετωπίζουν ζητήματα

μεταδεδομένων και διαλειτουργικότητας τόσο για τον ανθρωπο-χρήστη όσο και για τη μηχανική επεξεργασία. Εξετάζονται πιθανά σενάρια συντακτικών και σημασιολογικών μετασχηματισμών με σκοπό γλωσσικοί πόροι και υπηρεσίες να γίνουν προσβάσιμοι ομοιογενώς στη διεθνή ερευνητική κοινότητα, ξεπερνώντας τους περιορισμούς που προκύπτουν από την έλλειψη διαλειτουργικότητας σε όλα τα επίπεδα.

Τελικό στόχο όλων αυτών των δράσεων αποτελεί ουσιαστικά η μελλοντική δυνατότητα να αποκριθούμε θετικά για την αλήθεια του οράματος του Tim Berners Lee, διατυπωμένη στο Weaving the Web : "The vision I have for the Web is about anything being potentially connected with anything. It is a vision that provides us with new freedom, and allows us to grow faster than we ever could. . . . it brings the workings of society closer to the workings of our minds."

Στέλιος Πιπερίδης

Ηλεκτρολόγος Μηχανικός
Ινστιτούτο Επεξεργασίας Λόγου
Αρτέμιδος 6 & Επιδαύρου, 15125 Μαρούσι
Τηλ : 210-6875421
Ηλ. Δ/νση : spip@ilsp.gr

European Language Resources Association
Rue Brillat-Savarin 55-57, 75013 Paris