ΑΥΤΟΜΑΤΗ ΜΕΤΑΓΓΛΩΤΙΣΗ/ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ

Η ΕΜΠΕΙΡΙΑ ΤΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΚΡΗΤΗΣ

ΤΟΥ **ΓΑΝΝΗ ΚΟΣΜΑ,** της ομάδας μηχανογράφησης βιβλιοθηκών του Υπολογιστικού Κέντρου του Πανεπιστημίου Κρήτης/ΙΤΕ

Η εισήγησή μας αναφέρεται στην έρευνα πάνω σε θέματα μεταγραμματισμού στην ελληνική γλώσσα και ειδικότερα στην εμπειρία που έχει αποκτήσει το Πανεπιστήμιο Κρήτης μέσω του προγράμματος Helen. Ειδικότερα θα επιχειρηθεί:

- να δοθεί μια εικόνα της συνεισφοράς του Πανεπιστημίου Κρήτης στην μέχρι τώρα πορεία του Helen.

- να περιγραφεί η παρούσα φάση του προγράμματος μαζί με τα σχέδια και τις προγραμματισμένες δραστηριότητες για το άμεσο μέλλον.

- να γίνει αναφορά σε σημεία για πιθανή επέκταση του προγράμματος στο μέλλον, καθώς και σε ορισμένες προτάσεις για περαιτέρω μελέτη, που μπορεί να γίνει ακόμα και μετά την εξάντληση των χρονικών περιθωρίων του Helen.

Η συνεισφορά του Πανεπιστημίου Κρήτης στο HELEN.

Στην πρώτη φάση του Helen, κύριος σκοπός του ΠΚ ήταν κατ' αρχήν να αποκτήσει μια εμπειρία σχετικά με τα υπάρχοντα σχήματα μεταγραμματισμού. Συνεπώς έγινε μια διεξοδική μελέτη ώστε να κατανοηθεί ο τρόπος με τον οποίο αντιμετωπίζει το κάθε σχήμα την μετατροπή των χαρακτήρων καθώς και να ανιχνεύσει τις ιδιομορφίες που παρουσιάζονται στο μεταγραμματισμό ορισμένων γραμμάτων σε διάφορες περιπτώσεις. Έτσι πετύχαμε να έχουμε μια πολύ καλή εικόνα των σχημάτων μεταγραμματισμού και της χρήσης τους.

Σ' αυτό το χρονικό διάστημα του Πανεπιστήμιο Κρήτης ήταν ο κύριος υπεύθυνος των δοκιμών του λογισμικού το οποίο αναπτύσσονταν στο Πανεπιστήμιο του Bradford, σκοπός του οποίου ήταν η κατασκευής ενός πυρήνα με διαδικασίες μετατροπής χαρακτήρων ακολουθώντας κάποιο από τα σχήματα. Διεξήχθηκαν δοκιμές στο εν λόγω λογισμικό με στόχο αφ' ενός τον έλεγχο της ορθότητάς του, αλλά και την εξοικείωση στον τρόπο με τον οποίο αντιμετωπίζονται οι ιδιαιτερότητες ή/και οι ασάφειες των σχημάτων αυτών κατά την εφαρμογή.

Η μέθοδος που ακολουθήσαμε κατά τον πειραματισμό του λογισμικού, ήταν ανεξάρτητη από το υπολογιστικό περιβάλλον (PC, Mac, Unix), στο οποίο έγινε, καθώς και από το σχήμα μεταγραμματισμού το οποίο δοκιμάζαμε. Κατ' αρχήν έγινε μια επιλογή συγκεκριμένων ελληνικών λέξεων, οι οποίες κύριο χαρακτηριστικό τους έχουν ορισμένες ιδιαιτερότητες που παρουσιάζουν κάποια γράμματα ή συνδυασμοί γραμμάτων κατά τη γραφή τους, βάσει κάποιου σχήματος μεταγραμματισμού. Να σημειωθεί ότι καμμιά από τις λέξεις που δοκιμάσαμε δεν ήταν όνομα ή άλλη λέξη που είναι αντικείμενο διαφορετικού χειρισμού από το λογισμικό μεταγλώττισης. Δημιουργήθηκαν έτσι, σύμφωνα με καθένα από τα σχήματα μεταγραμματισμού, τμήματα κειμένων, που περιέχουν τις επιλεγμένες μαζί με πλήθος άλλων απλών λέξεων. Τέλος εφαρμόστηκε το λογισμικό σε καθένα από αυτά τα κείμενα στην απλή και στην αναλυτική (verbose) μορφή. Η αναλυτική μορφή σε περίπτωση ασαφειών, δίνει όλες τις πιθανές εναλλακτικές λύσεις για τη μετατροπή ενός γράμματος.

Θα πρέπει να σημειωθεί ότι στην περίπτωση του σχήματος μεταγραμματισμού ISO R843, έγινε δοκιμή και σε ένα μεγάλο αριθμό βιβλιογραφικών εγγραφών που είχαν παραληφθεί από την βιβλιοθήκη του King's College.

Έπειτα από μελέτη των αποτελεσμάτων της παράπάνω διαδικασίας, καταλήξαμε στο γενικό συμπέρασμα ότι οι ασάφειες που φαίνονται σε κάποια από τα αποτελέσματα, οφείλονται σε αδυναμίες των σχημάτων να χειριστούν κάποιες ιδιαιτερότητες και όχι σε δυσλειτουργίες των προγραμμάτων μεταγραμματισμού, τα οποία λειτουργούν ικανοποιητικά δίνοντας τα αναμενόμενα αποτελέσματα σε όλες τις περιπτώσεις. Οι αδυναμίες αυτές των σχημάτων μεταγλωττισμού, φαίνεται να επικεντρώνονται στην χρήση των τόνων και των διαλυτικών και πιο συγκεκριμένα σε όλα τα σχήματα εκτός των ISO R843 και Oxford. Επίσης φαίνεται να υπάρχει ασάφεια στη μετατροπή ορισμένων γραμμάτων. Σε ορισμένα σχήματα το ίδιο λατινικό γράμμα αντιστοιχεί σε περισσότερα από ένα ελληνικά.

Αυτό το οποίο είναι επίσης αξιοσημείωτο και αποτελεί αντικείμενο περαιτέρω προβληματισμού, είναι

η αδυναμία αυτόματης αναγνώρισης και εξαίρεσης ξενόγλωσσων λέξεων ή φράσεων από τη διαδικασία μεταγραμματισμού. Αυτό είναι φυσιολογικό, αφού ολόκληρο το κείμενο είναι γραμμένο με λατινικούς χαρακτήρες και δεν υπάρχει κάτι που να διακρίνει μονοσήμαντα τις ελληνικές από τις ξενόγλωσσες λέξεις.

Παραδείγματα αυτών των περιπτώσεων φαίνονται στους επόμενους πίνακες;

Παράδειγμα μεταγραμματισμού MARC εγγραφής

100	10	@a About, @b Edmond @c 1828-1885
245	12	@aO basile'us t on or'e on@d'Edmont Ampo'u
260	00	@a Ath ena@b Galax'ias @c 1968
300	00	@a 199p
440	00	@a Biblioth ek e El! en on kai X'en on Suggraf'e on @ 197
100	10	@Aa Αβουτ, @b Εδομνδ@c 1828-1885
245	12	@a Ο βασιλεύε των ορέων @d'Εντμοντ Αμπού
260	00	@a Αθήνα @b Γαλαξίας @c 1968
300	00	@a 199p
440	00	Βιβλιοθήκη Ελλήνων και Ξένων Συγγραφέων@v197

Παράδειγμα μεταγλωττισμού

Σχήμα	Είσοδος	Αναμενόμενο	Αποτέλεσμα
BL	g'ata	γάτα	γ/γκ ατα
tayt'oteta	ταυτότητα	ταυ ο/ω	τ ε/ τα
ΕΛΟΤ	b'ala	μπάλα	β..μπ αλα
	an'agki	ανάγκη	αναγκ ι/η
'	angelos	άγγελος	αγγελ ο/ως
IT	xana" ypolog'izo	ξαναϋπολογίζω	ξαναυπολογίζω
	pro''ypothes e	προϋπόθεση	προυποθεση
LC	b'ala	μπάλα	β/μπ αλα
Oxford	d'ala	μπάλα	μβ/μπ άλα
RAK	eykair'ia	ευκαιρία	ευκαιρια
	eyro''ik'os	ευρωπαϊκός	ευρωπαικος
Verbose	yalino	γιάλινο	

Τέλος θα πρέπει να σημειώσουμε την αξία που είχε το αυθεντικό κείμενο που είχαμε στη διάθεσή μας από τις βιβλιογραφικές εγγραφές της βιβλιοθήκης του King's Collge για τη διεξαγωγή των πειραμάτων μας για την περίπτωση του σχήματος ISO R 843. Η δυνατότητα δοκιμής με πραγματικά δεδομένα προερχόμενα από φόρμες που χρησιμοποιούν τα συγκεκριμένα σχήματα στην καθημερινή τους απασχόληση, δίνει μια άλλα διάσταση στην αξιοπιστία των αποτελεσμάτων του πειράματος. Αυτό οδηγεί στο συμπέρασμα ότι είναι σκόπιμο παραπέρα δοκιμές του λογισμικού να γίνονται με χρήση βιβλιογραφικών εγγράφων ή απλού κειμένου, τα οποία να προέρχονται από τις πηγές των σχημάτων μεταγραμματισμού.

Προβληματισμοί και προτάσεις

Από τα όσα εκθέσαμε μέχρι τώρα προκύπτουν κάποιοι προβληματισμοί, τους οποίους θα επιχειρήσομε να παραθέσουμε μαζί με προτάσεις για κάποια πιθανή λύση ή τους τρόπους προσέγγισης των προβλημάτων αυτών. Πρώτη παρατήρησή μας είναι ότι θα ήταν ευχής έργο να μην υπάρχει περιορισμός αλλά να μπορούν να συνυπάρχουν ελληνικές και ξένες λέξεις σε ένα κείμενο με λατινικούς χαρακτήρες, χωρίς οι λέξεις αυτές να αλλοιώνονται κατά περίπτωση από το λογισμικό μεταγλώττισης. Για το συγκεκριμένο πρόβλημα μια καλή λύση, η οποία εφαρμόστηκε σε μια έκδοση του προγράμματος μεταγλώττισης με επιτυχία, είναι να εσωκλείεται το ελληνικό κομμάτι κειμένου με κάποιες χαρακτηριστικές φράσεις, γνωστές από πριν σ' αυτό το κομμάτι κειμένου αφήνοντας τις ξενόγλωσσες φράσεις ανέπαφες. Αυτό μπορεί να εφαρμοστεί, με κάποια παραλλαγή ίσως, και στην περίπτωση των MARC εγγραφών. Εδώ θα μπορούμε να εσωκλείουμε το ελληνικό κείμενο με ειδικούς χαρακτήρες αντί για χαρακτηριστικές φράσεις για μεγαλύτερη ευκολία αλλά και εξοικονόμηση αποθηκευτικού χώρου. Βέβαια αυτή η λύση μπορεί να εφαρμοστεί σε κείμενα ή εγγραφές, που θα δημιουργούνται από τώρα και στο εξής. Για τα ήδη υπάρχοντα κείμενα και εγγραφές ίσως θα πρέπει να γίνει προσπάθεια προσαρμογής τους.

Σχετικά με την αντιμετώπιση αδυναμιών των σχημάτων μεταγραμματισμού, θα ήταν επιθυμητή η δυνατότητα ελέγχου της ορθότητας του κειμένου που προκύπτει από τη μεταγλώττιση με τη μικρότερη δυνατή, σε κόπο και χρόνο, παρέμβαση του τελικού χρήστη. Το πρόβλημα αυτό ίσως αντιμετωπίζεται σε μεγάλο βαθμό με τη χρήση ειδικού λογισμικού ελέγχο και διόρθωσης ορθογραφίας κειμένου. Τέτοιο λογισμικό έχει τη δυνατότητα διαγνωσης ανορθόγραφων λέξεων και διόρθωσης ή αντικατάστασής τους με παρεμφερείς.

Είδαμε ότι τα κύρια ονόματα αποτελούν μια ξεχωριστή κατηγορία λέξεων, ώς προς την αντιμετώπισή τους από το λογισμικό μεταγλώττισης. Κάποιο ερώτημα που δημιουργείται άμεσα είναι αν υπάρχουν κι

άλλες λέξεις που να τυγχάνουν τέτοιας μεταχείρισης. Συναφές μ' αυτό είναι οι περιπτώσεις λέξεων, οι οποίες θα πρέπει να είναι δεκτικές μετάφρασης και όχι μεταγλώττισης, όταν βρίσκονται σε κάποιο κείμενο. Αναφερόμαστε σε ορισμένες λέξεις που σχετίζονται με την ορολογία κάποιου κλάδου, όπως για παράδειγμα η λέξη computer που συναντάται πολλές φορές σε ελληνικά κείμενα, ή σύμβολα που πολλές φορές χρησιμοποιούνται με την ξενόγλωσση γραφή.

Μια επόμενη παρατήρηση αναφέρεται στη μορφή της αναλυτικής (verbose) εκτέλεσης του λογισμικού. Η χρήση που έγινε κατά την περίοδο της δοκιμής ήταν σχετικά περιορισμένη και εξυπηρετούσε επαρκώς τις ανάγκες μας. Όμως η συγκεκριμένη μορφή ίσως να μην είναι τόσο βολική σ' ένα πραγματικά περιβάλλον εφαρμογής. Ίσως η άμεση επιλογή κάποιας πιθανής προτεινόμενης λύσης από το χρήστη σε συνδυασμό με τη χρήση τεχνικών για τη δημιουργία μιας a priori γνώσης για τις μετέπειτα εφαρμογές του λογισμικού να είναι η πιο καλή αντιμετώπιση για την αντικατάσταση της αναλυτικής μορφής όπως την είδαμε μέχρι τώρα.

Τέλος θα θέλαμε να κάνουμε μια γενική παρατήρηση πάνω στα σχήματα μεταγραμματισμού. Καθένα σίγουρα εξυπηρετεί κάποιες τοπικές ανάγκες και δημιουργήθηκε για να προσφέρει λύσεις σε κάποια συγκεκριμένα προβλήματα, όμως, από τις εφαρμογές μεταγλωττισμού που κάναμε στα πλαίσια του HELEN , φαίνεται ότι ορισμένα από αυτά δεν προσφέρονται ως ιδεώδεις λύσεις για τέτοιους είδους χρήσεις. Θα πρέπει να τονίσουμε τη σαφήνεια που παρουσιάζει το ISO R 843 σχήμα μεταγραμματισμού. Αξίζει να σημειωθεί ότι είναι το μοναδικό που δεν παρουσίασε ούτε μια περίπτωση ασάφειας. Αντίστοιχα, κείμενα γραμμένα σύμφωνα με το σχήμα ΕΛΟΤ 743 είναι περισσότερο ευανάγνωστα.

Προγραμματισμένες δραστηριότητες

Με τη μέχρι τώρα ασχολία, τον πειραματισμό και τον προβληματισμό πάνω στα σχήματα μεταγραμματισμού, τα προβλήματα που παρουσιάζουν αλλά και τις δυνατότητες που δημιουργούν, έχει δημιουργηθεί μια πολλή καλή υποδομή για την ανάπτυξη ενός ευρύτερου προβληματισμού που θα προσφέρει ικανοποιητικές λύσεις σε θέματα εφαρμογής του μηχανισμού αυτόματου μεταγραμματισμού και της αποκατάστασης του αρχικού κειμένου στον ειδικό τομέα της αυτοματοποίησης λειτουργιών μιας βιβλιοθήκης.

Είναι γνωστό ότι υπάρχει πληθώρα βιβλιογραφικών εγγραφών καταλογογραφημένων σε βιβλιοθήκες του εξωτερικού. Προφανώς τα πεδία των εγγραφών είναι διαμορφωμένα σε λατινική γραφή ακολουθώντας κάποιο σχήμα μεταγραμαμτισμού. Καταλαβαίνουμε ότι η απόκτηση και η ικανότητα χρήσης τέτοιων εγγραφών από τις ελληνικές βιβλιοθήκες είναι άμεσου ενδιαφέροντος. Είναι λοιπόν προφανής η ανάγκη προσαρμογής του λογισμικού μεταγραμματισμού και κατάλληλης εφαρμογής του σ' ένα σύστημα μηχανοργάνωσης βιβλιοθήκης. Για την επιτυχία μιας τέτοιας προσέγγισης, χρειάζεται συλλογή πλήθους εγγραφών από βιβλιοθήκες του εξωτερικού, ώστε να μπορεί να γίνει μια στατιστική μελέτη και ανάλυση στο MARC. Με αυτόν τον τρόπο θα εξαχθούν συμπεράσματα και θα καταγραφούν προδιαγραφές για τη δημιουργία καταλλήλου λογισμικού που θα μπορεί να αντιμετωπίσει το ζήτημα μεταγραμματισμού των βιβλιογραφικών εγγραφών κατά τον καλύτερο δυνατόν τρόπο.

Από την άλλη πλευρά πρέπει να μελετηθούν οι τομείς του συστήματος μηχανοργάνωσης βιβλιοθήκης που μπορούν να συνεργαστούν με ένα τέτοιο εργαλείο. Η καθημερινή χρήση και λειτουργία του συστήματος υπαγορέυουν τις βασικές εφαρμογές του λογισμικού μεταγραμματισμού. Οι διαδικασίες μαζικής εισαγωγής και εξαγωγής βιβλιογραφικών εγγραφών είναι προφανές ότι διεκδικούν την μεγαλύτερη προτεραιότητα για προσαρμογή του λογισμικού.

Δύο άλλοι τομείς όπου μπορούν να αποτελέσουν πεδία εφαρμογής για τα προγράμματα αυτόματου μεταγλωττισμού, είναι ο αυτόματος κατάλογος πρόσβασης κοινού και η διαδικασία εισαγωγής και διόρθωσης εγγραφών (καταλογογράφηση). Ιδιαίτερα θα πρέπει να σταθούμε στην περίπτωση του αυτοματοποιημένου καταλόγου. Σήμερα με την ανάπτυξη των τηλεπικοινωνιών και την δυκτιακή υποδομή που έχει δημιουργηθεί, όλο και περισσότερες βιβλιοθήκες διαφόρων ιδρυμάτων και οργανισμών διασυνδέονται παγκοσμίως. Έτσι δίνεται η δυνατότητα σε χρήστες εκτός Ελλάδος να έχουν

πρόσβαση σε ελληνικούς αυτοματοποιημένους καταλόγους και να μπορούν να αναζητούν ελληνικές εγγραφές. Όμως σε πολλές περιπτώσεις είναι αδύνατη η απεικόνιση ή η εισαγωγή ελληνικών χαρακτήρων στον απομακρυσμένο τερματικό σταθμό, με αποτέλεσμα οι ελληνικοί κατάλογοι να μην είναι αναγνώσιμοι για τους χρήστες του εξωτερικού. Με την εφαρμογή του λογισμικού μεταγραμματισμού, αυτό το πρόβλημα είναι δυνατό να επιλυθεί.

Μια άλλη επιλογή η οποία μας απασχολεί είναι ο τρόπος φύλαξης των βιβλιογραφικών εγγραφών στο σύστημα αυτοματοποίησης σε συνδυασμό με τον χρόνο εφαρμογής του λογισμικού μεταγραμματισμού. Εδώ υπάρχουν δύο εναλλακτικές λύσεις:

- στη μία περίπτωση μπορούμε να επιλέξουμε τη διατήρηση της αρχικής εγγραφής με λατινικούς χαρακτήρες, μορφή. Εδώ η διαδικασία μεταγραμματισμού θα πρέπει να εφαρμόζεται κάθε φορά που θα ζητείται η απεικόνιση ή χρήση κάθε τέτοιας εγγραφής.

- εναλλακτικά μπορεί να επιλεγεί η φύλαξη των εγγραφών στη μεταγλωττισμένη, με ελληνικούς χαρακτήρες, μορφή. Στην περίπτωση αυτή η διαδικασία μεταγραμματισμού γίνεται μόνο μια φορά κατά την εισαγωγή της κάθε εγγραφής στο σύστημα αυτοματοποίησης.

Ίσως το σημείο στο οποίο η δεύτερη προσέγγιση υπερτερεί της πρώτης είναι το γεγονός ότι τα αποτελέσματα του μεταγραμματισμού είναι μόνιμα και έτσι μπορούν να διορθωθούν τυχόντα λάθη ή ασάφειες που σημειώθηκαν κατά την αρχική καταλογογράφηση από τον ξένο καταλογογράφο κι έτσι να επιτύχουμε καλύτερη ποιότητα εγγραφών. Κάθε μια από τις αναφερόμενες περιπτώσεις είναι αποδεκτή και είναι επιλογή της κάθε βιβλιοθήκης ποιά θα εφαρμόσει.

Μελλοντικές κατευθύνσεις

Θα θέλαμε, τέλος, να διατυπώσουμε κάποιες ενδιαφέρουσες προτάσεις για τη βελτίωση της απόδοσης στο μέλλον των συστημάτων στο μέλλον αυτόματης μεταγλώττισης βιβλιογραφικών αναγραφών και μεταφοράς τους από το ένα σύστημα στο άλλο.

Διαφαίνεται καθαρά η ανάγκη για την καθιέρωση ενός αμφιδρομικού και απόλυτα αντιστρέψιμου σχήματος μεταγραμματισμού προκειμένου να χρησιμοποιείται σε αυτοματοποιημένα συστήματα, για ανταλλαγή, απεικόνιση και ανεύρεση βιβλιογραφικών εγγραφών με τα κατά το δυνατόν καλύτερα αποτελέσματα.

Στο μέλλον θα πρέπει να γίνει επίσημη πρόταση σε κάποιο αρμόδιο φορέα για μια τροποποίηση του MARC, για υποστήριξη μικτού ελληνικού και ξενόγλωσσου κειμένου καθώς και τρόπου αναγνώρισης/διάκρισης του σχήματος μεταγραμματισμού που έχει εφαρμοστεί σε μια MARC εγγραφή.

Απαραίτητη κρίνεται η δημιουργία μιας διαδικασίας μετατροπής ελληνικού κειμένου σε μορφή με λατινικούς χαρακτήρες ακολουθώντας κάποιο από τα υπάρχοντα σχήματα μεταγραμματισμού, ώστε να αξιοποιηθούν οι χιλιάδες εκλατινισμένες βιβλιογραφικές αναγραφές που βρίσκονται στις βιβλιοθήκες του δυτικού κόσμου.

Για μια ολοκληρωμένη προσέγγιση του προβλήματος είναι σκόπιμη η προσπάθεια συγκέντρωσης κι άλλων σχημάτων μεταγραμματισμού ελληνικού κειμένου, πέραν αυτών που χρησιμοποιήθηκαν μέχρι σήμερα από το HELEN, που πιθανόν χρησιμοποιούνται από βιβλιοθήκες που φιλοξενούν ελληνόγλωσσο υλικό.

Θα πρέπει επίσης να επισημάνουμε ότι χρησιμοποιώντας την εμπειρία που έχει συγκεντρωθεί, είναι εφικτή η διεύρυνση του αυτόματου μεταγραμματισμού σε γλώσσες με παρόμοιες ανάγκες.

Συμπεράσματα

Καταλήγοντας θα πρέπει να τονίσουμε ότι για να αποφέρει τελικά καρπούς η προσπάθεια αυτή, χρειάζεται η συμμετοχή όλων των ενδιαφερομένων. Η συμβολή όλων στον προσδιορισμό αναγκών βάσει των καθημερινών τους λειτουργιών αλλά και στην επίλυση επιμέρους προβλημάτων θα είναι πολύ χρήσιμο. Με αυτόν τον τρόπο θα επιτύχουμε να βάλουμε μια αρχή και να δημιουργήσουμε τις προϋποθέσεις για ανταλλαγή ελληνικών και βιβλιογραφικών εγγραφών αλλά και άλλων ελληνικών κειμένων με φορείς του εξωτερικού. Επίσης θα δοθεί η δυνατότητα πρόσβασης των ελληνικών αυτοματοποιημένων καταλόγων από χρήστες του εξωτερικού και ευκαιρία προσέγγισης της ελληνικής βιβλιογραφίας και του ελληνικού γραπτού λόγου.