Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας στα Συστήματα Αναζήτησης των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών

Άννα Μάστορα[1], Μανόλης Πεπονάκης[2], Σαράντος Καπιδάκης[1]
[1]Εργαστήριο Ψηφιακών Βιβλιοθηκών και Ηλεκτρονικής Δημοσίευσης, Τμήμα Αρχειονομίας και Βιβλιοθηκονομίας, Ιόνιο Πανεπιστήμιο
{mastora, sarantos}@ionio.gr
[2]Εθνικό Κέντρο Τεκμηρίωσης, Εθνικό Ίδρυμα Ερευνών
epepo@ekt.gr

Περίληψη

Η αναντιστοιχία μεταξύ των όρων της ερώτησης που υποβάλλει ο χρήστης και των όρων που έχουν ευρετηριαστεί είναι ένα σημαντικό πρόβλημα, το οποίο επηρεάζει την ανάκτηση σχετικών τεκμηρίων κατά την αναζήτηση πληροφοριών. Σε γλώσσες με έντονη μορφολογία, όπως είναι η Ελληνική γλώσσα, η λέξη παίρνει διαφορετικές μορφές για να εκφράσει αριθμούς, πτώσεις, γένη, χρόνους κτλ. Δυσχεραίνεται, συνεπώς, το έργο της αναζήτησης πληροφοριών καθώς πρέπει να είναι εκ των προτέρων γνωστή η μορφή της λέξης που έχει ευρετηριαστεί ώστε να υποβληθεί ομοίως στο σύστημα κατά την αναζήτηση. Ένας τρόπος αντιμετώπισης του προβλήματος της έντονης μορφολογίας των γλωσσών, στο πλαίσιο της αναντιστοιχίας μεταξύ των υποβαλλόμενων όρων σε μια ερώτηση και εκείνων που περιλαμβάνονται στο αντίστοιχο ευρετήριο, σχετίζεται με τις Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας (ΕΤΓΕ). Στόχος της παρούσας μελέτης είναι να καταδείξει τα πλεονεκτήματα των Εφαρμογών Τεχνολογιών Γλωσσικής Επεξεργασίας στα συστήματα αναζήτησης των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών και ταυτόχρονα να καταδείξει τα προβλήματα που προκύπτουν όταν υλοποιούνται με διαφορετικό τρόπο οι παραπάνω τεχνολογίες. Για το λόγο αυτό, εξετάσαμε ως προς την εφαρμογή αυτών των τεχνολογιών Ελληνικές Ακαδημαϊκές Βιβλιοθήκες που διαθέτουν και OPAC και Ιδρυματικό Αποθετήριο καθώς και συγκρίναμε τη συνέπεια εφαρμογής αυτών τόσο μεταξύ των συστημάτων του ίδιου ιδρύματος όσο και οριζόντια, δηλαδή των ιδρυμάτων μεταξύ τους για να διαπιστώσουμε το βαθμό διαλειτουργικότητας που επιτυγχάνεται.

Λέξεις κλειδιά:

Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας, Πληροφοριακά Συστήματα, Δημόσιοι Κατάλογοι Βιβλιοθηκών, Αποθετήρια, Αναζήτηση Πληροφοριών

Abstract

A way of dealing with the problem of highly inflectional languages as well as the query – document terms mismatch problem is by implementing Language Processing Techniques. The aim of this study is to report, through presenting empirical data, on the Language Processing Techniques and their advantages if implemented by the information retrieval systems of Greek Academic Libraries. The objectives of this study are twofold. First goal is to acknowledge these techniques and then try to designate the interoperability issues deriving from the varying implementations. For this purpose, we examined Greek Academic Libraries which host both an OPAC and an Institutional Repository towards whether they implement any kind of language technology. Additionally, we report whether these techniques are implemented consistently in terms of the OPAC and the Institutional Repository of the same institution, as well as among different institutions.

Keywords:

Language Processing Techniques, Information Systems, Libraries' OPACs, Repositories, Information Search

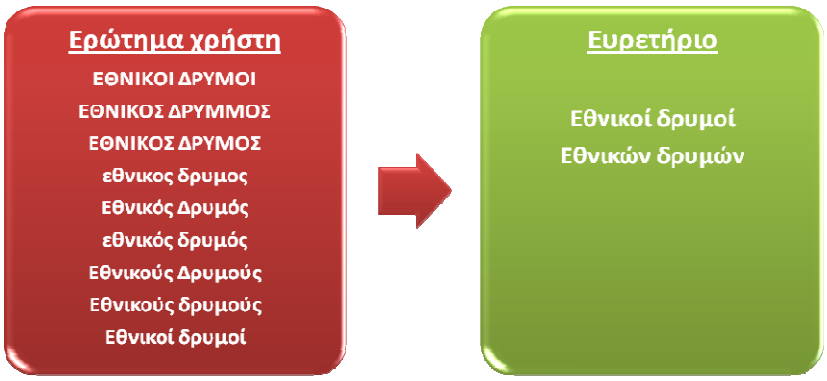
1. Εισαγωγή

Σε γλώσσες με έντονη μορφολογία, όπως είναι η Ελληνική γλώσσα, η λέξη παίρνει διαφορετικές μορφές για να εκφράσει αριθμούς, πτώσεις, γένη, χρόνους κτλ. και οι χρήστες αποδεικνύονται περισσότερο από πρόθυμοι να αξιοποιήσουν τις δυνατότητες που τους παρέχει η Ελληνική γλώσσα στη συνδιαλλαγή τους με τα συστήματα αναζήτησης πληροφοριών (Mastora and Kapidakis, 2012). Ωστόσο, αυτό το φαινόμενο, παρά τις δυνατότητες που παρέχει για καλύτερη έκφραση του ζητούμενου από την πλευρά του χρήστη, δεν έχει θετικές συνέπειες κατά την αναζήτηση και ανάκτηση πληροφοριών.

Στο σύγχρονο ψηφιακό περιβάλλον γίνεται περισσότερο αντιληπτή η επίδραση του φαινομένου καθώς βαίνει αυξανόμενη η χρήση λέξεων - κλειδιών για την αναζήτηση ακολουθώντας το παράδειγμα και τις πρακτικές των μηχανών αναζήτησης που αναζητούν περιεχόμενο στο Διαδίκτυο (Hearst, 2009). Συνεπώς φαίνεται ότι οι χρήστες αποφεύγουν να χρησιμοποιούν για τις αναζητήσεις τους ταξινομικούς αριθμούς αλλά και τη φυλλομέτρηση (browsing) καθιερωμένων λεξιλογίων (π.χ. θεματικών επικεφαλίδων) (Salaba, 2009). Άλλωστε η χρήση των ελεγχόμενων λεξιλογίων βαίνει μειούμενη στις ελληνικές

ακαδημαϊκές βιβλιοθήκες, αφού όπως έχει δείξει παλαιότερη έρευνα (Πεπονάκης και Σφακάκης, 2008), σε μεγάλο ποσοστό αποθετηρίων δεν χρησιμοποιείται ελεγχόμενο λεξιλόγιο στα πεδία των θεμάτων. Αυτή την πρακτική μεταφέρουν οι χρήστες και στα συστήματα αναζήτησης των βιβλιοθηκών, παρά το γεγονός ότι παρέχεται από αυτά δομημένη και σημασιολογικά προσδιορισμένη πληροφορία προστιθέμενης αξίας η οποία δίνει τη δυνατότητα εκτέλεσης πιο στοχευμένων αναζητήσεων.

Στο Σχήμα 1 αποτυπώνονται, με τη χρήση ενός παραδείγματος[1], οι διαφορετικοί τρόποι υποβολής ενός όρου από διαφορετικούς χρήστες και από την άλλη οι όροι που είναι διαθέσιμοι στη βάση δεδομένων. Είναι εμφανής η ποικιλομορφία καθώς παρατηρείται χρήση ενικού και πληθυντικού αριθμού, τονισμένων και άτονων χαρακτήρων, διαφορετικών πτώσεων, αλλά και ορθογραφημένων και ανορθόγραφων λέξεων. Στόχος κατά τη διαδικασία ανάκτησης πληροφοριών είναι –τουλάχιστον– η αντιστοιχία των δύο στηλών σε λεξικολογικό επίπεδο. Η τρέχουσα πρακτική απαιτεί από τον χρήστη να γνωρίζει εκ των προτέρων τη μορφή με την οποία έχει ευρετηριαστεί ο επιθυμητός όρος στη βάση δεδομένων ή να πετύχει την αντιστοιχία μέσω διαδοχικών δοκιμών και συνδυασμού κριτηρίων αναζήτησης. Αυτή η διαδικασία επιβαρύνει τον χρήστη κατά τη διαδικασία αναζήτησης και δεν έχει πάντα το επιθυμητό αποτέλεσμα.



Σχήμα 1: *Παράδειγμα υποβαλλόμενων ερωτήσεων και διαθεσιμότητα όρων βάσης δεδομένων*

Ωστόσο, υπάρχουν τεχνικές που δίνουν τη δυνατότητα αντιμετώπισης αυτού του φαινομένου χωρίς να απαιτείται ειδική γνώση από τον χρήστη ούτε να επιβαρύνεται με επιπλέον βήματα κατά την αναζήτηση πληροφοριών. Πρόκειται για τις Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας (ΕΤΓΕ), οι οποίες μπορούν να εφαρμοστούν σε μεγάλο βαθμό και στις περιπτώσεις αναζητήσεων με λέξεις - κλειδιά.

Η παρούσα μελέτη περιλαμβάνει στη συνέχεια τις εξής ενότητες. Στην ενότητα 2 γίνεται διατύπωση της ερευνητικής υπόθεσης και των αντικειμενικών σκοπών της έρευνας. Στην ενότητα 3 γίνεται σύντομη εισαγωγή στις Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας και σύνδεσή τους με τη διαδικασία της Ανάκτησης Πληροφοριών, ενώ στην ενότητα 4

[1] Το παράδειγμα προέρχεται από διαθέσιμα πειραματικά δεδομένα (Mastora and Kapidakis, 2012).

ακολουθεί η παρουσίαση της μεθοδολογίας της έρευνάς μας. Κατόπιν, στην ενότητα 5 γίνεται παρουσίαση των αποτελεσμάτων μας και κλείνουμε στην ενότητα 6 με τα συμπεράσματα και ειδικότερες επισημάνσεις για την εφαρμογή των ΕΤΓΕ, με ιδιαίτερη αναφορά σε θέματα διαλειτουργικότητας.

2. Υπόθεση έρευνας: στόχος και αντικειμενικοί σκοποί

Στόχος της παρούσας μελέτης είναι να καταδείξει τα πλεονεκτήματα από την εφαρμογή Τεχνολογιών Γλωσσικής Επεξεργασίας στα συστήματα αναζήτησης και ταυτόχρονα να καταδείξει τα προβλήματα που προκύπτουν όταν υλοποιούνται με διαφορετικό τρόπο οι παραπάνω τεχνολογίες. Για το λόγο αυτό οι αντικειμενικοί σκοποί της έρευνας διακρίνονται σε δύο μέρη. Πρώτα, βασιστήκαμε σε πειραματικά δεδομένα που προέκυψαν από την εφαρμογή εργαλείων γλωσσικής τεχνολογίας, τα οποία αναπτύξαμε στο πλαίσιο της ευρύτερης έρευνάς μας. Κατόπιν, εξετάσαμε Ελληνικές Ακαδημαϊκές Βιβλιοθήκες που διαθέτουν και OPAC και Ιδρυματικό Αποθετήριο σχετικά με την εφαρμογή Τεχνολογιών Γλωσσικής Επεξεργασίας. Στη συνέχεια, συγκρίναμε τη συνέπεια υλοποίησης των ΕΤΓΕ τόσο μεταξύ των συστημάτων του ίδιου ιδρύματος όσο και οριζόντια, δηλαδή των ιδρυμάτων μεταξύ τους. Ζητούμενο ήταν να διαπιστώσουμε το βαθμό διαλειτουργικότητας στην εφαρμογή των εν λόγω τεχνολογιών ώστε να επισημάνουμε τα προβλήματα που δημιουργούνται στο χρήστη και σε υπηρεσίες προστιθέμενης αξίας, οι οποίες βασίζονται στη διαλειτουργικότητα, από τη μη προβλεπόμενη εφαρμογή των τεχνολογιών.

3. Γλωσσικές τεχνολογίες: εισαγωγή

Η Ελληνική γλώσσα είναι μία γλώσσα πλούσια, με έντονη μορφολογία. Όπως περιγράφεται στο «The Greek Language in the Digital Age» (Gavrilidou et al, 2012) τα Ελληνικά είναι μια γλώσσα με πλούσιο κλιτικό σύστημα και το λεξιλόγιό της χαρακτηρίζεται από έκταση και μήκος λέξεων καθώς και όγκο. Μια αιτία για τον όγκο του λεξιλογίου είναι ο μεγάλος αριθμός συνωνύμων που παρατηρείται. Όπως συμβαίνει με όλες τις γλώσσες, το λεξιλόγιο περιλαμβάνει επίσης λέξεις δανεισμένες από άλλες γλώσσες. Κατά συνέπεια, για την ίδια έννοια είναι πιθανό να υπάρχουν 3 ή 4 λέξεις, καθεμιά προερχόμενη από μια διαφορετική γλώσσα. Μια άλλη αιτία για το εκτενές λεξιλόγιο είναι ο πλούτος του παραγωγικού μορφολογικού συστήματος: η παραγωγική αλυσίδα *ρήμα* > *ρηματικό ουσιαστικό* > *ονοματικό επίθετο* > *επίρρημα* είναι πολύ συνηθισμένη (π.χ. δημιουργώ > δημιουργία/δημιουργός > δημιουργικός > δημιουργικά). Επίσης, τα Ελληνικά χαρακτηρίζονται από ισχυρό παραγωγικό μηχανισμό για τα υποκοριστικά και τα μεγεθυντικά ουσιαστικών και επιθέτων.

Ένα επιπλέον πρόβλημα που δημιουργείται από τα χαρακτηριστικά της Ελληνικής γλώσσας έχει να κάνει με τον τρόπο υποβολής των ερωτημάτων στα συστήματα αναζήτησης. Έχει γίνει φανερό από προηγούμενες έρευνες στο συγκεκριμένο πεδίο (Lazarinis, 2007˙ Efthimiadis et al., 2009˙ Mastora and Kapidakis, 2012) ότι οι χρήστες αξιοποιούν τις

δυνατότητες της Ελληνικής γλώσσας για να εκφράσουν την πληροφοριακή τους ανάγκη, κάτι που οδηγεί σε ποικιλομορφία του υποβαλλόμενου ερωτήματος.

Ένας τρόπος αντιμετώπισης του προβλήματος της έντονης μορφολογίας των γλωσσών στην αναζήτηση και ανάκτηση πληροφοριών αλλά και της αναντιστοιχίας μεταξύ των υποβαλλόμενων όρων σε μια ερώτηση και εκείνων που περιλαμβάνονται στο αντίστοιχο ευρετήριο είναι οι Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας. Οι γλωσσικές τεχνολογίες στοχεύουν στην αυτόματη ανάλυση (και ίσως, κατανόηση) και παραγωγή γραπτών ή προφορικών εκφράσεων της φυσικής γλώσσας (Liddy, 1997) και ποικίλλουν σχετικά με την πολυπλοκότητα εφαρμογής τους και το πρόβλημα που καλούνται να αντιμετωπίσουν. Μερικά από τα πιο βασικά πεδία εφαρμογής γλωσσικών τεχνολογιών είναι η διόρθωση ορθογραφικών λαθών, η υποστήριξη συγγραφής κειμένου, η εκμάθηση γλώσσας υποβοηθούμενη από υπολογιστή, η ανάκτηση πληροφορίας, η εξαγωγή πληροφορίας, η αυτόματη περίληψη κειμένου, η απάντηση ερωτημάτων, η αναγνώριση και η σύνθεση φωνής (Gavrilidou et al, 2012).

Ειδικότερα, για το πεδίο της ανάκτησης πληροφορίας αποτελεί συνήθη πρακτική η εφαρμογή γλωσσικών τεχνολογιών όπως η αποκατάληξη (stemming), η λημματοποίηση (lemmatization), ο διαχωρισμός των λέξεων (tokenization), η διαχείριση σημείων στίξης και ανεπιθύμητων λέξεων, ο ορθογραφικός έλεγχος, ο εντοπισμός συνωνύμων, η μορφολογική και συντακτική ανάλυση καθώς και η αναγνώριση ονοματικών οντοτήτων (named entities). Επίσης, σε αυτές τις τεχνολογίες εντάσσονται και θέματα διαχείρισης των κεφαλαίων – πεζών (μικρών) χαρακτήρων καθώς και, ειδικά στην περίπτωση της Ελληνικής γλώσσας, το θέμα των τονούμενων και άτονων χαρακτήρων.

4. Μεθοδολογία

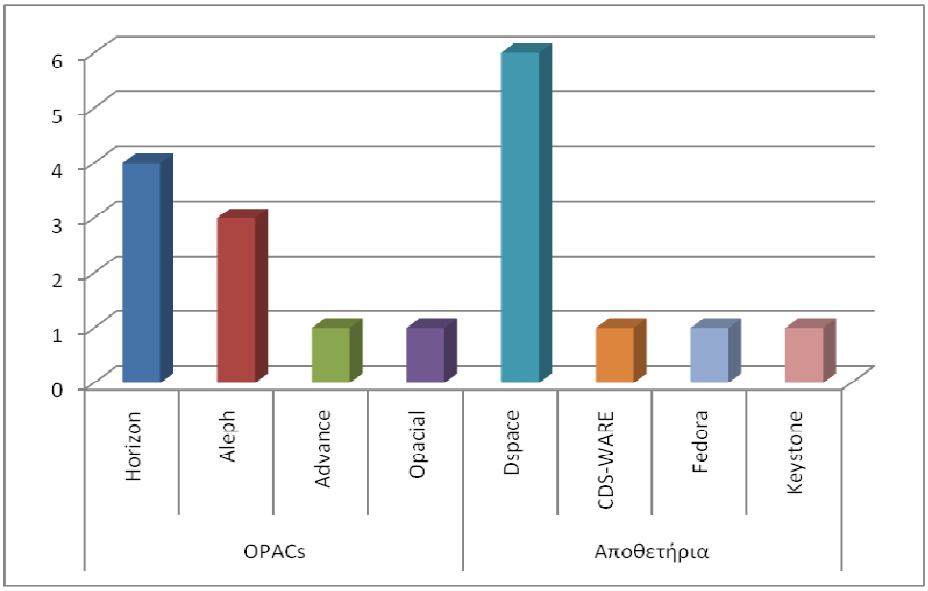
Χρησιμοποιήσαμε πειραματικά δεδομένα που προέκυψαν από σχετική έρευνα στο παρελθόν (Kapidakis et al, 2012). Τα αποτελέσματα αυτή της έρευνας κατέδειξαν τη διαφορετικότητα των μορφών με τις οποίες υποβάλλουν οι χρήστες τα ερωτήματά τους στα συστήματα αναζήτησης. Με δεδομένη την έντονη μορφολογία της Ελληνικής γλώσσας, το πρόβλημα της αναντιστοιχίας μεταξύ του υποβαλλόμενου ερωτήματος και των όρων της βάσης ήταν ιδιαίτερα έντονο. Στην προσπάθεια αντιμετώπισης του προβλήματος αναπτύξαμε μια σουίτα εργαλείων η οποία έχει τη δυνατότητα εφαρμογής ad-hoc τεχνολογιών γλωσσικής επεξεργασίας (π.χ. δυνατότητα απομάκρυνσης των stop words και διαχείρισης σημείων στίξης, μετατροπής των χαρακτήρων σε μικρούς ή κεφαλαίους καθώς και τονούμενους ή άτονους, καθώς και tokenization ή σύνθεση λέξεων) αλλά και δυνατότητα να καλέσει εξωτερικές εφαρμογές. Στην παρούσα φάση, χρησιμοποιείται για τον ορθογραφικό έλεγχο ο διορθωτής κειμένου Aspell[2] ενώ, για την αντιμετώπιση των πολλαπλών κλιτικών τύπων των

[2] Aspell, v.0.60.6. ©2000-2004 by Kevin Atkinson, http://aspell.net/ (Τελευταία πρόσβαση: Σεπτέμβριος 2012)

λέξεων εφαρμόσαμε λημματοποίηση[3] χρησιμοποιώντας τον λημματοποιητή (ilsp_nlp[4]) του Ινστιτούτου Επεξεργασίας Λόγου, ο οποίος διατίθεται ελεύθερα στο web.

Για τους σκοπούς της έρευνας εξετάσαμε τα συστήματα αναζήτησης ελληνικών ακαδημαϊκών βιβλιοθηκών ώστε να καταγραφεί η κατάσταση σχετικά με την εφαρμογή Τεχνολογιών Γλωσσικής Επεξεργασίας. Επιλέξαμε εκείνα τα ιδρύματα που είχαν δημόσιο κατάλογο (OPAC) και διέθεταν επίσης Ιδρυματικό Αποθετήριο. Δεν επιλέχθηκαν ιδρύματα τα οποία είχαν κάποιο σύστημα αποθετηρίου το οποίο δεν ήταν Ιδρυματικό αλλά είχε χρησιμοποιηθεί για τη συλλογή και διάθεση οποιουδήποτε είδους υλικού. Συνεπώς, ο περιορισμός ήταν η βιβλιοθήκη να παρέχει τόσο OPAC όσο και Ιδρυματικό Αποθετήριο με τα μεταδεδομένα ελεύθερα στο διαδίκτυο. Δέκα βιβλιοθήκες πληρούσαν τα παραπάνω κριτήρια. Ωστόσο, μόνο οι εννέα είχαν σε λειτουργία τόσο OPAC όσο και Ιδρυματικό Αποθετήριο το διάστημα από την 1η έως και την 20η Σεπτεμβρίου 2012 και αυτές μόνον συμπεριελήφθησαν στην έρευνα.

Στο παρακάτω γράφημα φαίνεται η κατανομή των συστημάτων αναζήτησης στους OPACs και τα αποθετήρια που περιλαμβάνονται στην έρευνα. Επισημαίνεται πως η κατανομή των OPACs δεν είναι απόλυτα αντιπροσωπευτική των συστημάτων αναζήτησης των ελληνικών ακαδημαϊκών βιβλιοθηκών αλλά η επιλογή καθορίστηκε από την ύπαρξη αποθετηρίου. Το ενδιαφέρον της έρευνάς μας επικεντρώθηκε στο σύστημα με το οποίο έρχεται σε επαφή ο χρήστης και όχι το ολοκληρωμένο σύστημα αυτοματισμού της βιβλιοθήκης. Για αυτό το λόγο εμφανίζεται στο παρακάτω γράφημα και στους πίνακες το σύστημα Opacial παρότι δεν πρόκειται για το ολοκληρωμένο σύστημα αυτοματισμού της βιβλιοθήκης.



Γράφημα 1: Κατανομή των συστημάτων που διαχειρίζονται OPACs και αποθετήρια

[3] Λημματοποίηση είναι η αναγωγή ενός όρου στον πρώτο κλιτικό του τύπο. Πρόκειται επί της ουσίας για τη μορφή του όρου όπως συναντάται στα ερμηνευτικά λεξικά. Για παράδειγμα, η λημματοποίηση του όρου «εθνικών δασών» έχει ως αποτέλεσμα το «εθνικός δάσος».

[4] Λημματοποιητής ilsp_nlp, http://nlp.ilsp.gr/soaplab2-axis/ (Τελευταία πρόσβαση: Σεπτέμβριος 2012)

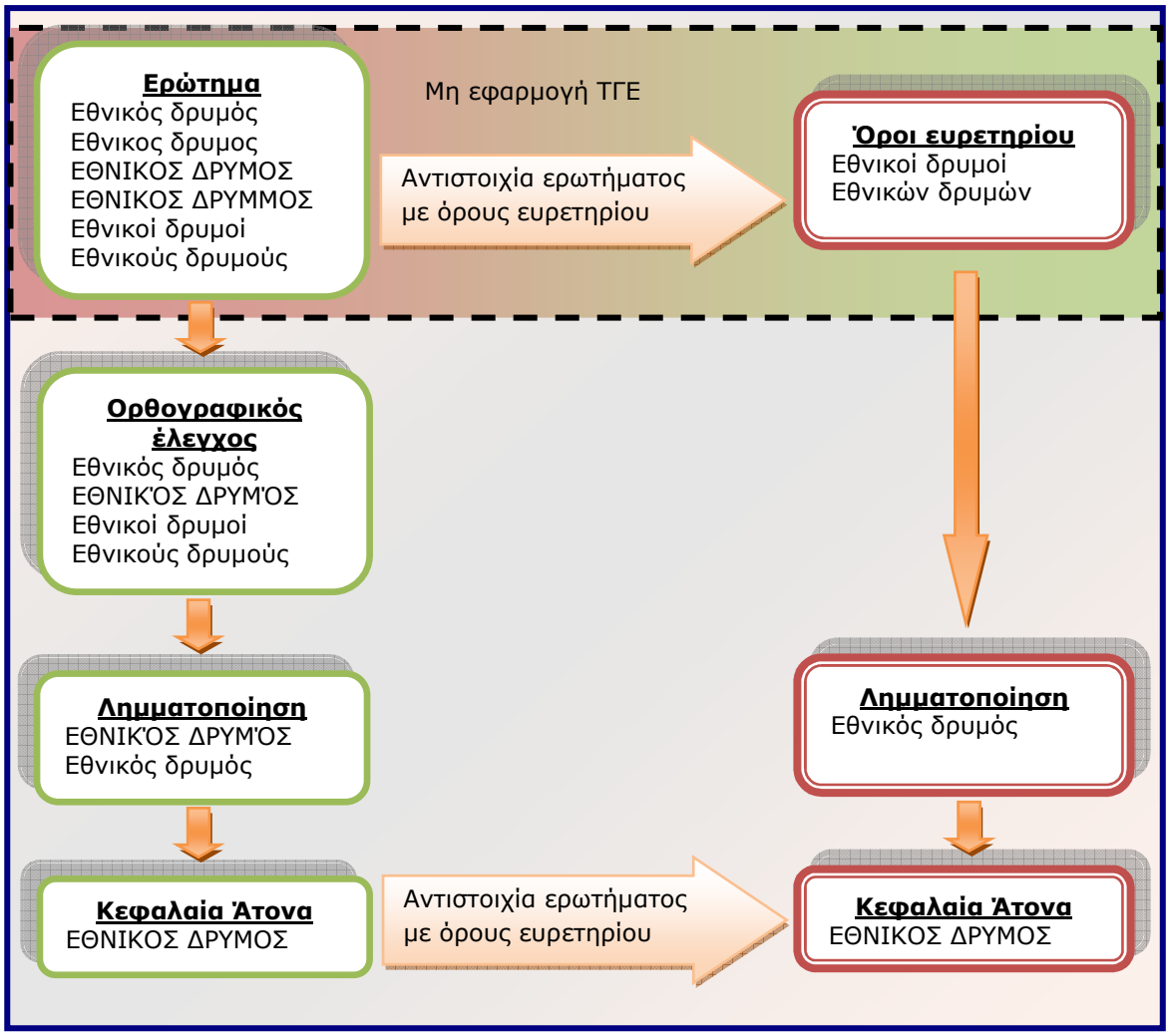
Οι αναζητήσεις υποβλήθηκαν στα συστήματα των βιβλιοθηκών μέσω του web interface που αυτά διέθεταν. Ο browser που χρησιμοποιήθηκε ήταν ο Firefox (v. 15.0.1) ενώ οι ερωτήσεις υποβλήθηκαν στο ευρετήριο «Θέμα» και στο ευρετήριο «Οποιοδήποτε». Τα σημεία που εξετάστηκαν ήταν σχετικά α) με τη διαχείριση των πεζών – κεφαλαίων χαρακτήρων β) με τη διαχείριση των τονούμενων – άτονων χαρακτήρων, γ) την ύπαρξη ορθογραφικού ελέγχου ή τη δυνατότητα προτεινόμενων όρων και δ) τη δυνατότητα εφαρμογής λημματοποίησης (lemmatization) ή αποκατάληξης (stemming). Για τις περιπτώσεις α και β ο έλεγχος γινόταν υποβάλλοντας ως ερώτηση μία λέξη με πεζούς και με κεφαλαίους χαρακτήρες και μία με τονούμενους και άτονους και ελεγχόταν εάν ο αριθμός των αποτελεσμάτων διέφερε αντίστοιχα. Αν διέφερε ο αριθμός των αποτελεσμάτων, εξεταζόταν εάν οι συγκεκριμένες εγγραφές που ανακτήθηκαν (οι οποίες διέθεταν τη λέξη γραμμένη με ένα συγκεκριμένο τρόπο) επιστρέφονταν όταν ο όρος αναζήτησης ήταν γραμμένος διαφορετικά. Για παράδειγμα, εάν η εγγραφή περιείχε τον όρο «Κρήτη» εξεταζόταν αν η εγγραφή αποτελούσε μέρος των αποτελεσμάτων όταν ο όρος αναζήτησης ήταν «Κρητη» (χωρίς τόνο). Για την περίπτωση γ υποβάλλονταν ερωτήματα τα οποία περιείχαν ανορθογραφίες, για παράδειγμα «Ιστωρία» ή «Κρύτη». Τέλος, για την περίπτωση δ υποβαλλόταν ένας όρος (ουσιαστικό) στον πρώτο κλιτικό τύπο (ονομαστική ενικού) και τα αποτελέσματα συγκρίνονταν με την αναζήτηση του ίδιου όρου γραμμένου στη γενική ενικού.

5. Αποτελέσματα

Σχετικά με τα πλεονεκτήματα από την εφαρμογή Τεχνικών Γλωσσικής Επεξεργασίας, διαπιστώσαμε ότι παρατηρείται σημαντική βελτίωση στον αριθμό των ερωτήσεων που επιστρέφουν αποτελέσματα μετά την εφαρμογή Τεχνικών Γλωσσικής Επεξεργασίας, σε σύγκριση με τον αριθμό των αποτελεσμάτων που επιστρέφεται χωρίς τη συνδρομή αυτών των τεχνολογιών. Πιο συγκεκριμένα, σε έρευνα που διενεργήσαμε με πειραματικά δεδομένα (Kapidakis et al., 2012), μετά την εφαρμογή του ορθογραφικού ελέγχου καταγράφηκε μείωση του ποσοστού των ερωτήσεων που επέστρεψαν μηδενικά αποτελέσματα κατά 9,75%. Ενώ μετά και την εφαρμογή της λημματοποίησης το ποσοστό των ερωτήσεων με μηδενικά αποτελέσματα έπεσε κατά 16,7% επί του συνόλου των αρχικά αποτυχημένων ερωτήσεων. Τα παραπάνω αποτελέσματα είναι ενδεικτικά και αφορούν μόνο στην επεξεργασία ερωτημάτων τα οποία είχαν αρχικά επιστρέψει μηδενικά αποτελέσματα.

Στο Σχήμα 2 αποτυπώνεται ένα **παράδειγμα** από την εφαρμογή επιλεγμένων Τεχνολογιών Γλωσσικής Επεξεργασίας. Στην κορυφή του σχήματος διακρίνουμε έξι διαφορετικούς τύπους με τους οποίους υποβλήθηκε το ερώτημα από τους χρήστες καθώς και δύο διαφορετικούς τύπους του όρου που υπάρχουν διαθέσιμοι στο ευρετήριο της βάσης δεδομένων. Στην περίπτωση που δεν εφαρμόζονται γλωσσικές τεχνολογίες, μόνο ένα από τα έξι υποβαλλόμενα ερωτήματα θα αντιστοιχιστεί με όρο από το ευρετήριο της βάσης και, μάλιστα, ακόμη και αυτός ο τρόπος θα οδηγήσει σε απώλεια σχετικών αποτελεσμάτων εφόσον ο δεύτερος όρος του ευρετηρίου δε θα εντοπιστεί. Αν, ωστόσο, ακολουθηθεί η διαδικασία ορθογραφικού ελέγχου, λημματοποίησης και μετατροπής των χαρακτήρων σε άτονων κεφαλαίων, τότε όλοι οι τύποι του υποβαλλόμενου ερωτήματος καταλήγουν στον

εξής έναν, και το ίδιο συμβαίνει για τους όρους του ευρετηρίου της βάσης. Διαπιστώνεται με αυτό τον τρόπο ότι τουλάχιστον σε λεξικολογικό επίπεδο υπάρχει ταύτιση των ερωτημάτων με τους όρους του ευρετηρίου και αποφεύγεται η απώλεια σχετικών αποτελεσμάτων.



Σχήμα 2: Παράδειγμα Εφαρμογής Τεχνολογιών Γλωσσικής Επεξεργασίας

Ακολούθως παρουσιάζονται τα αποτελέσματα που προέκυψαν από την έρευνα στα συστήματα αναζήτησης των βιβλιοθηκών. Όπως προαναφέρθηκε εξετάστηκαν δύο ευρετήρια: Θέμα και Οποιοδήποτε. Ωστόσο, επειδή ο τρόπος διαχείρισης του ερωτήματος από κάθε σύστημα ήταν πανομοιότυπος και για τα δύο ευρετήρια, οι τιμές στις στήλες των πινάκων είναι κοινές και για τα δύο.

Στον Πίνακα 1 παρουσιάζονται τα αποτελέσματα σχετικά με τα χαρακτηριστικά των OPACs που εξετάσαμε. Όπως φαίνεται υπάρχει απόλυτη συνέπεια μεταξύ των OPACs των βιβλιοθηκών στον τρόπο που χειρίζονται τα τονούμενα και τα άτονα, δηλαδή δεν διαφοροποιούν τα αποτελέσματα της αναζήτησης ανάλογα με το αν το υποβαλλόμενο ερώτημα περιέχει τόνους ή όχι. Στο σημείο αυτό είναι σημαντικό να επισημανθεί πως δεν αναφερόμαστε σε πολυτονικό σύστημα αλλά σε μονοτονικό της σύγχρονης δημοτικής γλώσσας. Σχετικά με τους προτεινόμενους όρους ή τη δυνατότητα ορθογραφικού ελέγχου, μόνον οι OPACs των βιβλιοθηκών Β, Γ και Θ προτείνουν κάποιους όρους όταν αποτύχει η αναζήτηση. Διαπιστώσαμε ότι αυτή η λειτουργία δεν αποτελεί ορθογραφικό έλεγχο, αλλά

όταν η αναζήτηση έχει μηδενικό αποτέλεσμα, προτείνονται όροι από το εν λόγω ευρετήριο οι οποίοι είναι αλφαβητικά «κοντά» προς τον υποβαλλόμενο όρο. Σε αυτή την περίπτωση, οι προτεινόμενοι όροι παραπέμπουν σε όρους που υπάρχουν στη βάση και οι οποίοι δεν είναι κατ' ανάγκη σωστά ορθογραφημένοι. Τέλος, διαπιστώνεται πως οι OPACs δεν εφαρμόζουν αποκατάληξη ή λημματοποίηση.

OPACs				
Βιβ/θήκη	Σύστημα	Διαφοροποιεί τονούμενα-άτονα	Ορθογραφικός έλεγχος ή προτεινόμενοι όροι	Αποκατάληξη ή λημματοποίηση
Α	Horizon	Όχι	Όχι	Όχι
Β	Aleph	Όχι	Ναι	Όχι
Γ	Aleph	Όχι	Ναι	Όχι
Δ	Horizon	Όχι	Όχι	Όχι
Ε	Advance	Όχι	Όχι	Όχι
ΣΤ	Horizon	Όχι	Όχι	Όχι
Ζ	Opacial	Όχι	Όχι	Όχι
Η	Horizon	Όχι	Όχι	Όχι
Θ	Aleph	Όχι	Ναι	Όχι

Πίνακας 1ος: *OPACs και Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας*

Η συνοχή του Πίνακα 1 ως προς τα τονούμενα και άτονα φαίνεται να διασπάται στα συστήματα των αποθετηρίων που παρουσιάζονται στον Πίνακα 2. Οι τρεις κατάλογοι αποθετηρίων στους εννιά δεν εξισώνουν τα τονούμενα με τα άτονα. Για τους προτεινόμενους όρους στον κατάλογο του αποθετηρίου της βιβλιοθήκης Α ισχύει ό,τι αναφέρθηκε στην προηγούμενη παράγραφο για τους OPACs των βιβλιοθηκών Β, Γ και Θ, δηλαδή όταν η αναζήτηση έχει μηδενικό αποτέλεσμα, προτείνονται όροι από το διαθέσιμο ευρετήριο οι οποίοι είναι αλφαβητικά «κοντά» προς τον υποβαλλόμενο όρο. Τέλος, διαπιστώνεται ότι τα συστήματα των αποθετηρίων δεν εφαρμόζουν αποκατάληξη ή λημματοποίηση.

Συστήματα Αποθετηρίων				
Βιβ/θηκη	Σύστημα	Διαφοροποιεί τονούμενα-άτονα	Ορθογραφικός έλεγχος ή προτεινόμενοι όροι	Αποκατάληξη ή λημματοποίηση
Α	CDS-WARE	Όχι	Ναι	Όχι
Β	DSpace	Όχι	Όχι	Όχι
Γ	Keystone	Όχι	Όχι	Όχι
Δ	DSpace	Ναι	Όχι	Όχι
Ε	DSpace	Ναι	Όχι	Όχι
ΣΤ	DSpace	Ναι	Όχι	Όχι
Ζ	Fedora	Όχι	Όχι	Όχι
Η	DSpace	Όχι	Όχι	Όχι
Θ	DSpace	Όχι	Όχι	Όχι

Πίνακας 2ος: *Ιδρυματικά Αποθετήρια και Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας*

Μια σημαντική παρατήρηση, η οποία δεν αποτυπώνεται στους πίνακες, έχει να κάνει με τη διαχείριση των κεφαλαίων και των μικρών (πεζών) χαρακτήρων. Ενώ, όπως αναφέρθηκε, τα συστήματα (τόσο των βιβλιοθηκών όσο και των αποθετηρίων) δεν επηρεάζονται από τα κεφαλαία ή πεζά, εξαίρεση αποτελεί ο χαρακτήρας που αντιστοιχεί στο τελικό σίγμα, «ς». Κατά κανόνα, ο χρήστης είτε γράψει με μικρά είτε με κεφαλαία το ερώτημά του θα πάρει το

ίδιο αποτέλεσμα. Ωστόσο, στις περιπτώσεις που η λέξη τελειώνει σε «ς» είναι πιθανό να υπάρξει πρόβλημα. Συγκεκριμένα το πρόβλημα προκύπτει από τη μη εξίσωση του «ς» με το «Σ» ή το «σ» και αυτό ισχύει για τους OPACs των βιβλιοθηκών Α, Δ, ΣΤ, Η καθώς και για τον κατάλογο του αποθετηρίου της βιβλιοθήκης Ε. Είναι προφανές πως το πρόβλημα είναι έντονο για τους χρήστες που πληκτρολογούν τον όρο αναζήτησης με κεφαλαία γράμματα.

6. Συμπεράσματα

Οι διαφορετικοί κλιτικοί τύποι ενός όρου μπορεί να επηρεάσουν σημαντικά την απόδοση της ανάκτησης πληροφορίας, ιδιαίτερα στις δικτυακές μηχανές αναζήτησης. Όταν πρόκειται για γλώσσες με έντονη μορφολογία («πλούσιο» κλιτικό σύστημα), η αποτυχία επεξεργασίας των διαφόρων μορφολογικών τύπων των λέξεων μπορεί να οδηγήσει μόνο σε 2%-10% επιτυχή απόδοση της ανάκτησης πληροφορίας (Ζώτος, 2007).

Ενώ εκ πρώτης όψης οι ΕΤΓΕ φαίνεται να είναι κάτι που μπορεί να βοηθήσει πολύ, δεν υπάρχει ουσιαστική χρήση τους. Στο Σχήμα 2 αποτυπώνεται η βελτίωση που θα μπορούσε να επιτευχθεί με την εφαρμογή ήπιων τεχνικών γλωσσικής επεξεργασίας. Εφαρμόζοντας γλωσσικές τεχνολογίες οι έξι διαφορετικοί τύποι του όρου που υπέβαλαν διαφορετικοί χρήστες κατά την αναζήτηση κατέληξαν στον εξής έναν. Αντίστοιχα, οι δύο διαφορετικοί τύποι του όρου που υπήρχαν στη βάση κατέληξαν στον εξής έναν με εφικτή πλέον την απόλυτη αντιστοιχία μεταξύ τους.

Με δεδομένη την ύπαρξη εργαλείων και τεχνικών που θα μπορούσαν να συμβάλουν αποφασιστικά στην ανάπτυξη των υπηρεσιών αναζήτησης που προσφέρουν οι βιβλιοθήκες είναι σημαντικό να διερευνηθούν οι δυνατότητες συνέργειας στο πλαίσιο μιας διεπιστημονικής προσέγγισης του θέματος. Όμως παρότι αυτά τα εργαλεία υπάρχουν και λειτουργούν σε πειραματικό στάδιο είναι αναγκαίο για τη βελτίωση των εργαλείων και την προσαρμογή τους στο περιβάλλον των βιβλιοθηκών να ελεγχθούν σε πραγματικές συνθήκες. Για το λόγο αυτό είναι αναγκαία η υποβολή τους σε πραγματικά δεδομένα ικανού αριθμού ώστε να αναδειχτούν με ακρίβεια τα πλεονεκτήματα αλλά και οι προκλήσεις που πρέπει να αντιμετωπιστούν.

Για παράδειγμα, πολλές φορές οι χρήστες επιθυμούν την ανάκτηση τεκμηρίων, στα οποία οι όροι να περιέχονται με τον τρόπο που είναι διατυπωμένοι στο ερώτημά τους καθώς σε κάποιες περιπτώσεις η διατήρηση του τονικού σημείου ή του κεφαλαίου πρώτου γράμματος μιας λέξης αποσαφηνίζει τη σημασία της. Τέτοιες περιπτώσεις αποτελούν τα «τονικά παρώνυμα», όπως *Αθήνα – Αθηνά*, *τσίπουρα – τσιπούρα*, *γέρος – γερός*, *νόμος – νομός*. Χαρακτηριστικό, επίσης, παράδειγμα, αποτελούν στη Νέα Ελληνική ορισμένα κύρια ονόματα, η συμβολοσειρά των οποίων ταυτίζεται απόλυτα με κάποιους κλιτικούς τύπους ουσιαστικών, π.χ. *Μαργαρίτα* (κύριο όνομα), *μαργαρίτα* (ουσιαστικό). Η μόνη διαφοροποίηση των δυο όρων σε τέτοιες περιπτώσεις έγκειται στον πρώτο χαρακτήρα, ο οποίος είναι κεφαλαίο αλφαβητικό σύμβολο στην περίπτωση των κυρίων ονομάτων και πεζό αλφαβητικό στην περίπτωση των ουσιαστικών (Ζώτος, 2007).

Αναφορικά με το δεύτερο αντικειμενικό σκοπό της παρούσας μελέτης, δηλαδή τη διερεύνηση υιοθέτησης Εφαρμογών Τεχνολογιών Γλωσσικής Επεξεργασίας από τα συστήματα αναζήτησης των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, επισημαίνουμε ότι αυτές είναι σημαντικά περιορισμένες σε σχέση με τις υπάρχουσες δυνατότητες. Περιλαμβάνουν ήπιες τεχνικές, κυρίως αναφορικά με τη διαχείριση τονούμενων και άτονων χαρακτήρων καθώς και τη διαχείριση κεφαλαίων και πεζών.

Διαπιστώνεται, επίσης, ότι ακόμη και στα σημεία όπου υπάρχει εφαρμογή βασικών γλωσσικών τεχνολογιών, οι διαφορετικές υλοποιήσεις δείχνουν να περιπλέκουν και όχι να απλοποιούν την κατάσταση. Για παράδειγμα, οι τρεις στους εννέα καταλόγους αποθετηρίων δεν εξισώνουν τα τονούμενα με τα άτονα. Τονίζεται πως αυτό δεν έχει να κάνει με διαφορετικά συστήματα αλλά, όπως προκύπτει από τον πίνακα 1, η διαφοροποίηση καταγράφεται σε υλοποιήσεις του ίδιου συστήματος (DSpace). Έτσι, από έξι συνολικά εγκαταστάσεις του DSpace στις τρεις τα τονούμενα δεν εξισώνονται με τα άτονα ενώ στις άλλες τρεις εξισώνονται.

Πέραν όμως της ασυνέπειας μεταξύ των διαφορετικών ιδρυμάτων εμφανίζεται σοβαρή ασυνέπεια μεταξύ των δύο καταλόγων του ίδιου ιδρύματος. Πάνω από τις μισές από τις βιβλιοθήκες (5/9) δεν ακολουθούν την ίδια υλοποίηση αναφορικά με τα θέματα που εξετάζονται και αποτυπώνονται στους Πίνακες 1 και 2. Επιπρόσθετα, αν ληφθεί υπόψη και η διαφοροποίηση ανάμεσα στο «Σ» και το «ς» τότε οι οχτώ στις εννιά βιβλιοθήκες δεν έχουν συνέπεια ανάμεσα στον OPAC και το αποθετήριο στα θέματα που εξετάζουμε. Δηλαδή ο χρήστης πρέπει να ακολουθήσει διαφορετική προσέγγιση στην αναζήτησή του ανάλογα με το αν ψάχνει στον OPAC ή τον κατάλογο του αποθετηρίου.

Στην εποχή της διάθεσης μεγάλου αριθμού δεδομένων και ύπαρξης μεγάλων συσσωρευτών (aggregators) περιεχομένου είναι πολύ σημαντική η χρήση κοινών πολιτικών ώστε να επιτυγχάνεται διαλειτουργικότητα. Όπως επισημαίνεται σε σχετική μελέτη (Koukourakis 2011, σ.111) αυτό που είναι ύψιστης σημασίας από εθνική σκοπιά, είναι η υιοθέτηση εθνικής πολιτικής για την Ανοικτή Πρόσβαση και τα ζητήματα διαλειτουργικότητας. Η παραπάνω επισήμανση έχει ακόμη πιο βαρύνουσα σημασία όταν πρόκειται για θέματα που σχετίζονται με τη γλώσσα ενώ πηγάζει από την πεποίθηση ότι για ένα θέμα που πρωτίστως αφορά μια συγκεκριμένη χώρα οι λύσεις δεν πρόκειται να προκύψουν από έρευνα εκτός των συνόρων. Οφείλουμε, όχι μόνο να συμμετέχουμε αλλά να είμαστε πρωτοπόροι σε θέματα που σχετίζονται με την Ελληνική γλώσσα καθώς ακόμη και για λύσεις που προέρχονται από διαφορετικές (αλλά παρόμοιας μορφολογίας) γλώσσες, θα πρέπει να γίνουν οι αναγκαίες παραμετροποιήσεις ώστε να υπάρξουν τα επιθυμητά αποτελέσματα και στην Ελληνική γλώσσα.

Σημείωση: *Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) – Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος ΙΙ. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.*

Βιβλιογραφία

Efthimiadis, E. et al., 2009. Non-english web search: an evaluation of indexing and searching the Greek web. *Information Retrieval*, 12(3), pp.352–379. Available at: http://dx.doi.org/10.1007/s10791-008-9084-6.

Gavrilidou, M., Koutsombogera, M. & Patrikakos, A., 2012. *The Greek Language in the Digital Age*, Springer. Available at: http://www.meta-net.eu/whitepapers/e-book/greek.pdf.

Hearst, M.A., 2009. *Search User Interfaces* 1st ed., Cambridge University Press. Available at: http://searchuserinterfaces.com/book/.

Kapidakis, S., Mastora, A. & Peponakis, M., 2012. Query Expansion of Zero-Hit Subject Searches: Using a Thesaurus in Conjunction with NLP Techniques. In P. Zaphiris et al., eds. *Theory and Practice of Digital Libraries*. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 433–438. Available at: http://dx.doi.org/10.1007/978-3-642-33290-6_48.

Koukourakis, M., 2011. Greek Academic Repositories: Policies for Making Available Scientific and Cultural Content. In *New Trends in Qualitive and Quantitative Methods in Libraries: Selected Papers Presented at the 2nd Qualitative and Quantitative Methods in Libraries: Proceedings of the International Conference on QQML2010 Chani*. New Trends in Qualitive and Quantitative Methods in Libraries. Athens: World Scientific Publishing, pp. 103–120.

Lazarinis, F., 2007. An initial exploration of the factors influencing retrieval of Web images in Greek queries. In *Proceedings of the 2007 Euro American conference on Telematics and information systems*. EATIS '07. New York, NY, USA: ACM, pp. 69:1–69:4. Available at: http://doi.acm.org/10.1145/1352694.1352765.

Liddy, E. D. (1998). Natural Language Processing for Information Retrieval and Knowledge Discovery. In *P. A. Cochrane, & E. H. Johnson (Eds.): Visualizing Subject Access for 21st Century Information Resources* [papers presented at the 1997 Clinic on Library Applications of Data Processing, March 2-4, 1997]: 137-147.

Mastora, A. & Kapidakis, S., 2012. Query Rewriting Using Shallow Language Processing: Effects on Keyword Subject Searches. In *International Workshop on Supporting User's Exploration on Digital Libraries*. pp. 3–14. Available at: http://ixa2.si.ehu.es/suedl/SUEDLproceedings.pdf.

Salaba, A., 2009. End-User Understanding of Indexing Language Information. *Cataloging Classification Quarterly*, 47(1), pp.23–51. Available at: http://dx.doi.org/10.1080/01639370802451983.

Ζώτος, Νικόλαος, 2007. *Εξατομικευμένη αναζήτηση πληροφορίας με χρήση σημασιολογικών δικτύων*. Μεταπτυχιακή Εργασία. Πανεπιστήμιο Πάτρας. Τμήμα Μηχανικών Η/Υ και Πληροφορικής. Διαθέσιμο στο: http://nemertes.lis.upatras.gr/jspui/handle/10889/642.

Πεπονάκης, Μανόλης & Σφακάκης, Μιχάλης, 2008. Αξιολόγηση διαλειτουργικότητας μεταδεδομένων μεταξύ των ιδρυματικών αποθετηρίων και των καταλόγων (OPACs) των Ελληνικών ακαδημαϊκών βιβλιοθηκών. Στο *17ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών: Η αξιολόγηση των Βιβλιοθηκών ως στοιχείο ποιότητας των Ακαδημαϊκών Ιδρυμάτων*. Ιωάννινα. Διαθέσιμο στο: http://eprints.rclis.org/handle/10760/13276.