Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας στα Συστήματα Αναζήτησης των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών

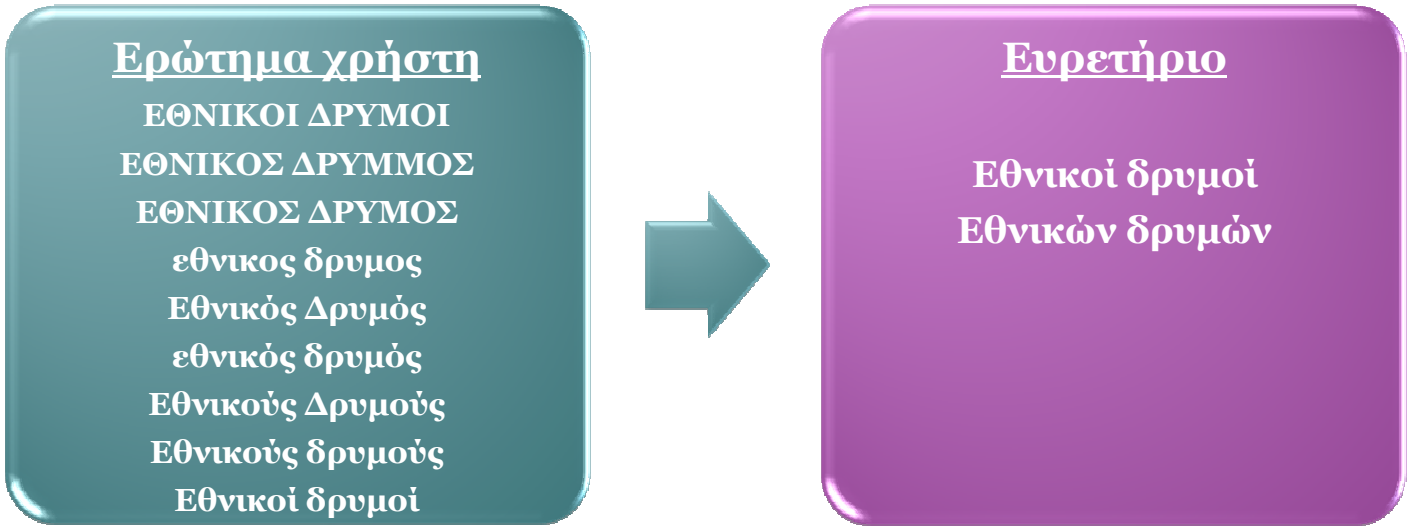
Άννα Μάστορα[1], Μανόλης Πεπονάκης[2], Σαράντος Καπιδάκης[1]

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

21ο Πανελλήνιο Συνέδριο Ακαδημαϊκών Βιβλιοθηκών
18-19 Οκτωβρίου 2012, Πειραιάς

Κίνητρο της έρευνας

- Έντονη μορφολογία ελληνικής γλώσσας
- Αυξημένη χρήση της φυσικής γλώσσας στη διατύπωση ερωτημάτων
- Περιορισμένη χρήση ελεγχόμενου λεξιλογίου
- Ποικιλομορφία στη διατύπωση του ερωτήματος από τους χρήστες



Γλωσσικές τεχνολογίες

- Στοχεύουν
  - Αυτόματη ανάλυση (και ίσως κατανόηση;) & παραγωγή γραπτών ή προφορικών εκφράσεων της φυσικής γλώσσας

- Πεδία εφαρμογής
  - Διόρθωση ορθογραφικών λαθών,
  - Εκμάθηση γλώσσας υποβοηθούμενη από Η/Υ,
  - Εξαγωγή πληροφορίας,
  - Αυτόματη περίληψη κειμένου,
  - Ανάκτηση πληροφορίας
  - κτλ.

Γλωσσικές τεχνολογίες: Ανάκτηση πληροφορίας

- Αποκατάληξη (stemming)
- Λημματοποίηση (lemmatisation)
- Διαχωρισμός λέξεων (tokenisation)
- Διαχείριση σημείων στίξης & ανεπιθύμητων λέξεων
- Ορθογραφικός έλεγχος
- Εντοπισμός συνωνύμων
- Μορφολογική & συντακτική ανάλυση
- Αναγνώριση ονοματικών οντοτήτων (Named Entity Recognition)
- Διαχείριση χαρακτήρων: κεφαλαίων – πεζών, τονούμενων – άτονων

Στόχος έρευνας

- Ανάδειξη πλεονεκτημάτων εφαρμογής γλωσσικών τεχνολογιών
  - Με έμφαση στην ελληνική γλώσσα
- Καταγραφή της υφιστάμενης κατάστασης αλλά και των προβλημάτων που προκύπτουν από την έλλειψη διαλειτουργικότητας στις εφαρμογές
  - Με έμφαση στα συστήματα αναζήτησης των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών

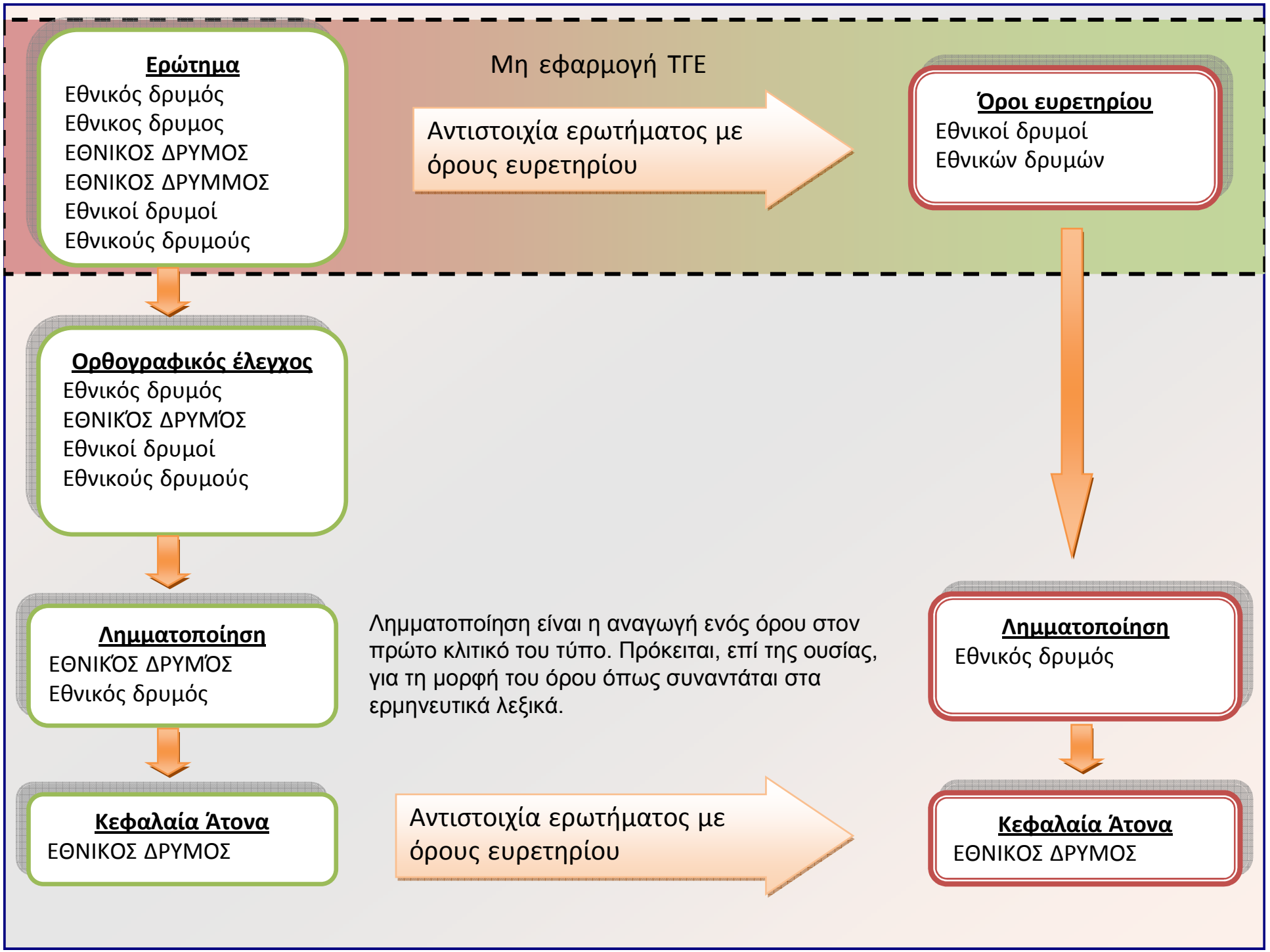
Αντικειμενικοί σκοποί

- Ανάδειξη αποτελεσμάτων από πειραματικά δεδομένα με εφαρμογή γλωσσικών τεχνολογιών
- Έρευνα σε Ελληνικές Ακαδημαϊκές Βιβλιοθήκες και Ιδρυματικά Αποθετήρια
  - Συγκριτική προσέγγιση ώστε να διαπιστωθεί η διαλειτουργικότητα

Δεδομένα & εργαλεία πειράματος

- Διαπιστώθηκε η χρήση διαφορετικών μορφολογικών τύπων της ελληνικής γλώσσας στην υποβολή των ερωτημάτων
- Διαπιστώθηκε αναντιστοιχία μεταξύ όρων ερωτήματος και όρων ευρετηρίου
- Αναπτύχθηκε σουίτα εργαλείων
  - Εφαρμογή ad-hoc τεχνολογιών γλωσσικής επεξεργασίας
    - Διαχείριση τονούμενων-άτονων, διαχείριση σημείων στίξης, μετατροπή χαρακτήρων σε πεζούς ή κεφαλαίους, κτλ
  - Δυνατότητα χρήσης εξωτερικών εργαλείων
    - Ορθογραφικός έλεγχος (ASpell) -> Βελτίωση* 9,75%
    - Λημματοποίηση (ilsp_nlp, από το ΙΕΛ) -> Βελτίωση* 16,7%

* Βελτίωση θεωρήθηκε η δυνατότητα ανάκτησης αποτελεσμάτων σε ερωτήματα που αρχικά είχαν μηδενικά αποτελέσματα


Ερώτημα
Εθνικός δρυμός
Εθνικος δρυμος
ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ
ΕΘΝΙΚΟΣ ΔΡΥΜΜΟΣ
Εθνικοί δρυμοί
Εθνικούς δρυμούς
Μη εφαρμογή ΤΓΕ
Αντιστοιχία ερωτήματος με όρους ευρετηρίου
Όροι ευρετηρίου
Εθνικοί δρυμοί
Εθνικών δρυμών
Ορθογραφικός έλεγχος
Εθνικός δρυμός
ΕΘΝΙΚΌΣ ΔΡΥΜΌΣ
Εθνικοί δρυμοί
Εθνικούς δρυμούς
Λημματοποίηση
ΕΘΝΙΚΌΣ ΔΡΥΜΌΣ
Εθνικός δρυμός
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.
Λημματοποίηση
Εθνικός δρυμός
Κεφαλαία Άτονα
ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ
Αντιστοιχία ερωτήματος με όρους ευρετηρίου
Κεφαλαία Άτονα
ΕΘΝΙΚΟΣ ΔΡΥΜΟΣ

Συμπεράσματα εφαρμογής ΤΓΕ

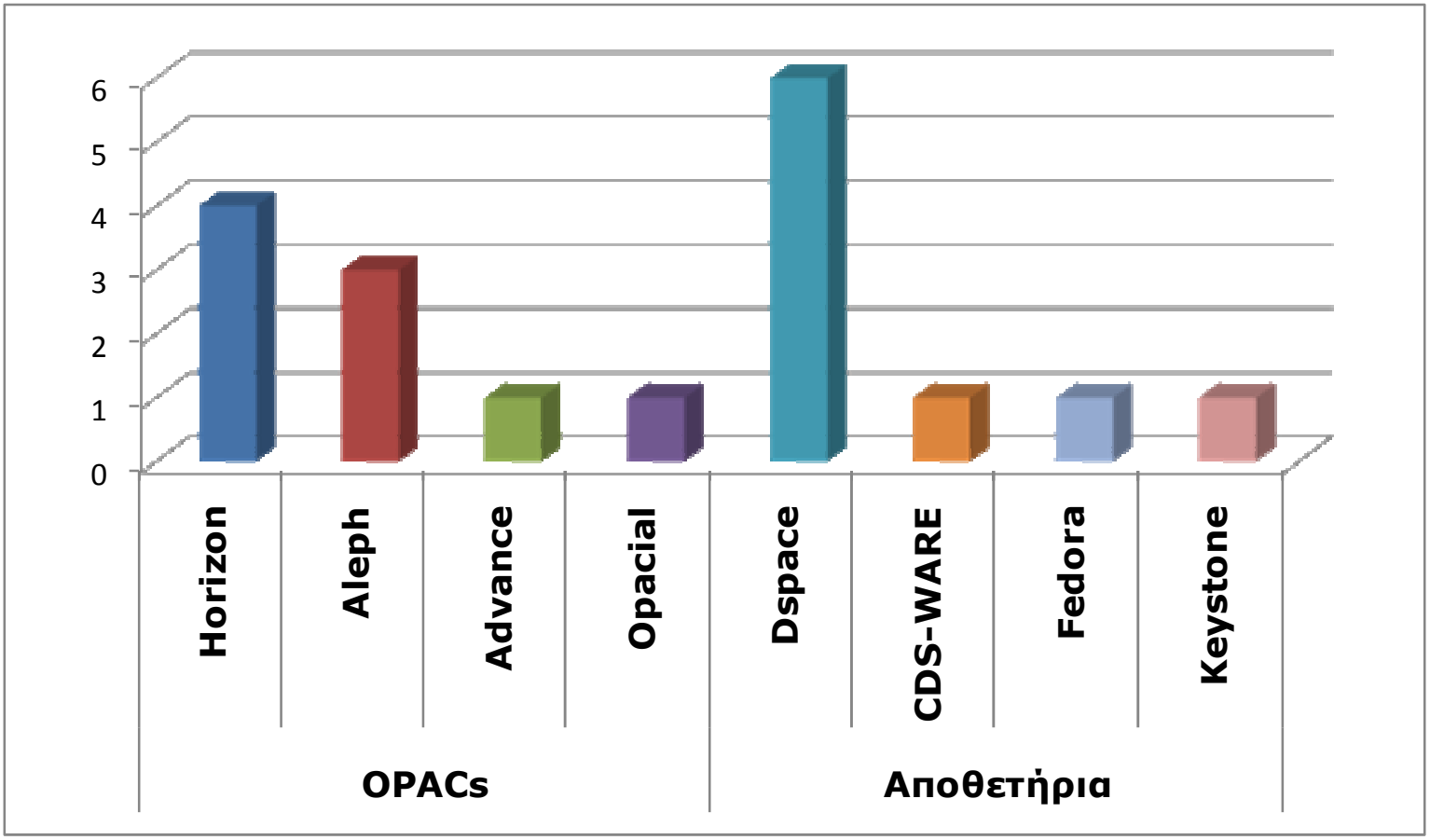
- Οι διαφορετικοί κλιτικοί τύποι ενός όρου μπορεί να επηρεάσουν σημαντικά την απόδοση της ανάκτησης πληροφορίας
- Εκ πρώτης όψης οι ΕΤΓΕ φαίνεται να είναι κάτι που μπορεί να βοηθήσει πολύ, ωστόσο, δεν υπάρχει ουσιαστική χρήση τους από τα συστήματα αναζήτησης ελληνικού περιεχομένου
- Δεδομένης της ύπαρξης εργαλείων και τεχνικών που θα μπορούσαν να συμβάλουν αποφασιστικά στην ανάπτυξη των υπηρεσιών αναζήτησης είναι σημαντικό να διερευνηθούν οι δυνατότητες συνέργιας στο πλαίσιο διεπιστημονικής προσέγγισης του θέματος

Οι γλωσσικές τεχνολογίες στα συστήματα αναζήτησης των Ακαδημαϊκών Βιβλιοθηκών

- Επιλέξαμε: εκείνα τα ιδρύματα που είχαν δημόσιο κατάλογο (OPAC) και διέθεταν επίσης Ιδρυματικό Αποθετήριο

- Εξετάσαμε: τα συστήματα αναζήτησης των παραπάνω ιδρυμάτων ώστε να καταγραφεί η κατάσταση σχετικά με την εφαρμογή Τεχνολογιών Γλωσσικής Επεξεργασίας

Κατανομή συστημάτων που εξετάστηκαν



OPACs και Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας

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

Ιδρυματικά Αποθετήρια και Εφαρμογές Τεχνολογιών Γλωσσικής Επεξεργασίας

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

Συμπεράσματα ΕΤΓΕ σε συστήματα Ελληνικών Ακαδημαϊκών Βιβλιοθηκών

- Τα συστήματα αναζήτησης των Ελληνικών Ακαδημαϊκών Βιβλιοθηκών εφαρμόζουν σε περιορισμένη έκταση γλωσσικές τεχνολογίες
  - τονούμενα-άτονα
    - διαφορετικές πολιτικές τόσο μεταξύ των συστημάτων κάθε ιδρύματος όσο και των ιδρυμάτων μεταξύ τους
  - κεφαλαία-πεζά
    - όλα τα συστήματα αναζήτησης (OPACs & Αποθετήρια) εξισώνουν κεφαλαία-πεζά
      - Εξαίρεση: το τελικό σίγμα «ς»: μη εξίσωση τού «ς» με το «Σ» ή το «σ» για τους OPACs των βιβλιοθηκών Α, Δ, ΣΤ, Η και στο Αποθετήριο Ε

Επισήμανση!

- Η εξίσωση κεφαλαίων-πεζών ή τονούμενων-άτονων χαρακτήρων δεν αποτελεί (πάντοτε) ζητούμενο (τονικά παρώνυμα κτλ.)
  - Αθήνα – Αθηνά
  - τσίπουρα – τσιπούρα
  - γέρος – γερός
  - νόμος - νομός
  - Μαργαρίτα – μαργαρίτα

Συμπεράσματα: σύνοψη

- Οι γλωσσικές τεχνολογίες μπορούν να έχουν θετικά αποτελέσματα για την ανάκτηση πληροφοριών από τα συστήματα αναζήτησης
  - Ώστε να διευκολύνουν το χρήστη στην αναζήτηση
    - Όπως έδειξαν τα αποτελέσματα της έρευνάς μας, η λημματοποίηση μπορεί να παίξει καθοριστικό ρόλο στην ανάκτηση ελληνικού περιεχομένου

- Είναι πολύ σημαντική η χρήση κοινών πολιτικών ώστε να επιτυγχάνεται διαλειτουργικότητα
  - Διαφορετικά, ο χρήστης βρίσκεται στην ίδια (δύσκολη) θέση όπως και χωρίς την εφαρμογή ΤΓΕ

Μελλοντική έρευνα

- Είναι αναγκαία η χρήση (δοκιμή) των εργαλείων και τεχνικών ΕΤΓΕ σε μεγάλο όγκο δεδομένων ελληνικού περιεχομένου
  - ❖Ανάγκη απόκτησης δεδομένων **και** από ελληνικούς φορείς για επιβεβαίωση των πειραματικών δεδομένων
- Δοκιμή περισσότερων εργαλείων και πιο εκλεπτυσμένων τεχνικών για πιο στοχευμένη ερμηνεία της πρόθεσης του χρήστη (query intent)
  - Μορφοσυντακτική ανάλυση
  - Αναγνώριση ονοματικών οντοτήτων
  - Εντοπισμός συνωνύμων, κτλ.
- Μελέτες συμπεριφοράς χρηστών για εξαγωγή-επιβεβαίωση μοντέλων αναζήτησης πληροφορίας

ΕΥΧΑΡΙΣΤΩ ΓΙΑ ΤΗΝ ΠΡΟΣΟΧΗ ΣΑΣ!

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.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης