



Πρόσβαση στο περιεχόμενο ιστορικών εγγράφων με χρήση εργαλείων Γλωσσικής Τεχνολογίας

Ελένη Γαλιώτου
Τμήμα Πληροφορικής, ΤΕΙ Αθήνας



- Η βιβλιοθήκη της Ι.Μ. Ευαγγελισμού της Θεοτόκου Σκιάθου – Ψηφιοποίηση
- Το ερευνητικό έργο «ΠΟΛΥΤΙΜΟ»
- Πρόσβαση στο περιεχόμενο των ιστορικών βιβλίων και χειρογράφων
- Μορφολογικός επεξεργαστής για την «πρώιμη» Νέα Ελληνική Γλώσσα
- Πρόσβαση στο σημασιολογικό περιεχόμενο
- Συμπεράσματα



Ι.Μ. Ευαγγελισμού Σκιάθου - Βιβλιοθήκη

- 1794 : Ίδρυση της Μονής από ομάδα μοναχών του κινήματος των «Κολλυβάδων».
 - Πρώτος ηγούμενος: Οσιος Νήφων
- Βιβλιοθήκη
 - Βιβλία διαφόρων λογίων μοναχών
 - Εντυπα βιβλία (16^{ου} – 19^{ου} αιω.)
 - Χειρόγραφα (11^{ου} – 19^{ου} αιω.)
- Έργα:
 - Πατριάρχου Φωτίου, Νικοδήμου του Αγιορείτου, Αθανασίου του Παρίου, κ.α.
 - Ομήρου, Θουκυδίδη
 - Γραμματικής, Μαθηματικών, Γεωγραφίας κ.α.



Ψηφιοποίηση της βιβλιοθήκης

- 30^{ος} Ηγούμενος της Μονής : Αρχιμ. Αγγελος Λύσσαρης
- Ευθύνη του Τμήματος Πληροφορικής του ΤΕΙ Αθήνας
- Στόχος:
 - διευκόλυνση της πρόσβασης στα βιβλία και χειρόγραφα
 - διαφύλαξη των βιβλίων και των χειρογράφων αυτά από τη συχνή χρήση
- ~ 250.000 σελίδες
- Ορισμένα βιβλία δίγλωσσα:
 - Ελληνικά – Λατινικά
- Έχουν εκδοθεί τα 3 πρώτα DVD (59 βιβλία)



«ΠΟΛΥΤΙΜΟ»

- Σύστημα επεξεργασίας, διαχείρισης και παροχής πρόσβασης στο περιεχόμενο πολύτιμων βιβλίων και χειρογράφων (ΓΓΕΤ: Επιχειρησιακό Πρόγραμμα «Κοινωνία της Πληροφορίας»).
- Συμετέχοντες φορείς:
 - BSI Α.Ε.
 - Εργαστήριο Υπολογιστικής Ευφυΐας ΕΚΕΦΕ «Δ»
 - Τμήμα Πληροφορικής ΤΕΙ Αθήνας
 - Συνεργάτιδες:
 - Αγγελική Ράλλη, Καθ. Γλωσσολογίας, Τμήμα Φιλολογίας Παν. Πατρών, Διευθύντρια Εργαστηρίου Νεοελληνικών Διαλέκτων
 - Ιωάννα Μανωλέσσου, Ερευνήτρια Β', Ακαδημία Αθηνών



Σώμα κειμένων

- 110 εικόνες που αντιστοιχούν σε σελίδες έντυπων βιβλίων (17^{ου} -18^{ου} αιω.)
 - Χρύσανθος (Νοταράς) *Ιστορία και Περιγραφή της Αγίας Γης και Αγίας Πόλεως Ιερουσαλήμ*, (1728)
 - 2. Ηλίας Μηνιάτης, *Διδαχαί* (1644)
 - 3. Μάξιμος (Μαργούνιος) *Βίοι Αγίων* (1685)
 - 4. Μελέτιος, *Γεωγραφία* (1728)
 - 5. Μελέτιος, *Εκκλησιαστική Ιστορία*, 1 τ. (1783)
 - 6. Ευστράτιος Αργέντης, *Σύνταγμα κατά Αζύμων* (1760)
 - 7. Ιωάσαφ Κορνήλιος, *Λόγοι Ηθικοί*, 1 τ. (1788)



Κριτήρια επιλογής κειμένων

- Γλώσσα η οποία να πλησιάζει την καθομιλουμένη/δημώδη της εποχής (άρα να προσφέρεται για γλωσσολογική ανάλυση)
- 2. Ομοειδής θεματική ώστε να μπορούν να χρησιμοποιηθούν οι ίδιες/παρόμοιες λέξεις – κλειδιά ως όροι αναζήτησης
- 3. Επαρκής αλλά όχι υπερβολικά μεγάλη έκταση κειμένου.



Πλαίσιο

- Εντοπισμός λέξεων απευθείας στο περιεχόμενο των ψηφιοποιημένων εγγράφων παρακάμπτοντας συμβατικές μεθόδους οπτικής αναγνώρισης χαρακτήρων που αποδεικνύονται αναποτελεσματικές στην περίπτωση των ιστορικών εγγράφων της Μονής. (Εργαστήριο Υπολογιστικής Ευφυΐας ΕΚΕΦΕ «Δ»)
- Χρήση προηγμένων τεχνικών επεξεργασίας φυσικής γλώσσας για την πραγματοποίηση «έξυπνων» αναζητήσεων κατά τη διάρκεια της διαδικασίας εντοπισμού των λέξεων



Τεχνικές Επεξεργασίας Φυσικής Γλώσσας

- Επέκταση, εμπλουτισμός των όρων αναζήτησης
- 1. Γεννήτρια λέξεων
 - Για τον εντοπισμό όλων των κλιτών μορφών μιας λέξης στο κείμενο
- 2. Λεξικό συνωνύμων
 - Πρόσβαση στο σημασιολογικό περιεχόμενο των εγγράφων
 - Εμπλουτισμό των αποτελεσμάτων της αναζήτησης.



Μορφολογική Γεννήτρια

- Αποτελεσματική αναζήτηση στα έγγραφα:
 - Λαμβάνονται υπ' όψιν όλες οι πιθανές κλιτές μορφές μιας λέξης χωρίς ρητή διατύπωση από τον χρήστη.
 - Χρήση συστήματος μορφολογικής γένεσης βασισμένο σε γλωσσολογική ανάλυση εάν οι όροι προέρχονται από μία γλώσσα με πλούσια μορφολογία (π.χ. Ελληνικά)
- Αναζήτηση σε ιστορικά έγγραφα
 - Σημαντικές διαφορές ανάμεσα στη γλώσσα του κειμένου και σύγχρονη μορφή της γλώσσας (ορθογραφία, μορφολογική αλλαγή κ.α.)



Μορφολογική Γεννήτρια (συν.)

- Μορφολογία της Πρώιμης Νέας Ελληνικής
 - Δεν έχει γίνει συστηματική ανάλυση –ακόμη.
 - Στοιχεία από την Αρχαία, Μεσαιωνική και Νέα Ελληνική
- Ονοματική κλίση
- Εμφαση στην επεκτασιμότητα του συστήματος
 - Εύκολος εμπλουτισμός με καινούργιες λέξεις με απλή αντιστοίχιση με τις λέξεις που είναι αντιπροσωπευτικές των κλιτικών τάξεων.
 - Εύκολη επέκταση του συστήματος και σε άλλα μορφολογικά φαινόμενα που μπορεί να αναλυθούν στο μέλλον.



Μορφολογική Γεννήτρια (συν.)

- Ανάπτυξη λογισμικού μορφολογικής επεξεργασίας
 - Ενσωματώνει τα εργαλεία SFST:
 - SFST :
 - Αναπτύχθηκε στο Ινστιτούτο Επεξεργασίας Φυσικής Γλώσσας του Παν. Στουτγάρδης .
 - Ενσωματώνει ένα μεταγλωττιστή που μεταφράζει κανονικές εκφράσεις σε μετατροπείς πεπερασμένων καταστάσεων που αναπαριστούν τη σχέση μεταξύ επιφανειακών μορφών γλώσσας με τις αντίστοιχες μορφές λεξικού.
- Κατά τη διάρκεια της γένεσης ο χρήστης μπορεί να:
 - εισάγει μία νέα λέξη-κλειδί
 - αναθέσει την κατάλληλη κλιτική τάξη επιλέγοντας τον εκπρόσωπο της τάξης αυτής



Παράδειγμα μορφολογικής γεννήτριας

Παράγωγα

Επιλεγμένη λέξη κλειδί: γλωσσα

Πρότυπα: θαλασσα

Παράγωγα της επιλεγμένης λέξης κλειδί

Παράγωγα

Λέξεις κλειδιά

- αζυμος
- διωγμος
- μυστηριον
- γλωσσα

γλωσσα
γλωσσαν
γλωσσας
γλωσσης
γλωσσων

Αποθήκευση

CEvent

Η παραγωγή των λέξεων ολοκληρώθηκε

OK



Πρόσβαση στο σημασιολογικό περιεχόμενο

- Στόχος : Διευκόλυνση της πρόσβασης στο σημασιολογικό περιεχόμενο των εγγράφων και εμπλουτισμός των αποτελεσμάτων τη αναζήτησης.
- Οι όροι εξάγονται συνήθως από:
 - Σημασιολογικά δίκτυα όπως το WordNet
 - Δίκτυα βασισμένα στις συν-εμφανίσεις
 - Υβριδική προσέγγιση
- Αποκλειστική χρήση σημασιολογικών δικτύων
 - Σημαντική βελτίωση στην απόδοση
- Οι όροι στα ιστορικά έγγραφα της συλλογής προέρχονται κυρίως από θρησκευτικό λεξιλόγιο.



Πρόσβαση στο σημασιολογικό περιεχόμενο (συν.)

- Σχέση συνωνυμίας προσαρμοσμένη σε ένα συγκεκριμένο θεματικό πεδίο (Turcato et al.)
 - Λειτουργικότητα αντίστοιχη με ένα σημασιολογικό δίκτυο
 - Αποφεύγονται ασάφειες που ίσως επηρεάζουν την ακρίβεια και την αποδοτικότητα του συστήματος.
- Ο χρήστης έχει τη δυνατότητα να:
 - προσθέσει, διορθώσει, διαγράψει μία λέξη με έως 5 συνώνυμα.
 - Κάθε συνώνυμο έχει μία σταθμισμένη τιμή συνάφειας με τη λέξη.
 - Οι τιμές κυμαίνονται από 1 έως το 10 (μεγαλύτερη συνάφεια ~ σταθμισμένη τιμή 1).
 - αναζητήσει λέξεων και συνωνύμων.
 - εξάγει όλες τις κλιτές μορφές μίας λέξης καθώς και τα σταθμισμένα συνώνυμα (με χρήση της κατάλληλης γραφικής διεπαφής χρήστη).



Εξαγωγή κλιτών μορφών και σταθμισμένων συνωνύμων

■ Διαχείριση παραγώγων - συνωνύμων

Επιλεγμένη Λέξη Κλειδί Παράγωγα της επιλεγμένης λέξης κλειδί Συνώνυμα της επιλεγμένης λέξης κλειδί

Λέξεις Κλειδιά	Παράγωγα	Συνώνυμα	Βάρος
▶ αυτοκρατωρ	▶ αυτοκρατορα	▶ βασιλευς	2
σφαλμα	αυτοκρατορας		
κακια	αυτοκρατορες		
πραξις	αυτοκρατορος		
ερωσ	αυτοκρατορων		
νεκρωσις			

Navigation icons: Home, Back, Forward, Stop, Play, Volume, Mute, Repeat, Refresh



Χρήση των κλιτών μορφών και συνωνύμων στη διαδικασία εντοπισμού λέξεων

Α' 152 Ν
3^ο 1

Ε' ΚΚΛΗΣΙΑΣΤΙΚΗ ΚΩ. Ε

και ὡς θέλωσι νὰ ἐισέλθωσιν εἰς τὴν Ἐκκλησίαν, ἢ ἑκατηγόρου δια' φέ-
νου δεσμῶ, πληγὰς ἀδικίας, καὶ τραύματα καὶ πυρκαϊκὰ τῶν
αὐτῶν ὅσων καὶ τὰς ὑπ' αὐτὸν Ἐπισκόπους, καὶ ἄλλας πολ-
λὰς.

§. 4. Ἐπειδὴ δὲ ὁ **Βασιλεὺς** καταφρόνησεν ὡς ψευδεῖς τὰς κατηγο-
ρίας αὐταῖν, ἔδωκεν ἀδειαν τῷ Ἀθανασίῳ νὰ ἐπιστρέψῃ εἰς τὴν ἐπαρχίαν
ταύτην, ἢ ἔγραψε πρὸς τὸν λαὸν τῆς Ἀλεξανδρείας νὰ μαρτυρήσωσι τὴν κα-
λωσύνην, ἢ τὴν οὐδὴν πίστιν. ἢ μὲ χαρὰν νὰ τὸν ὑποδεχθῶσι, ἢ νὰ
εἶναι πεπληροφορημένοι. ὡς ἐνάρετος ἄνθρωπος εἶναι, ἢ θεῖος, ἢ πῶς
φθόνῳ κινούμενοι οἱ αἰρετικοὶ ἔγραψαν κατ' αὐτῶν, ἢ πῶς ἦν ἀδῶος. Ἀπέ-
σας δὲ ὁ **Βασιλεὺς**, πῶς πολλοὶ ἔτι Αἰγύπτιοι, συγχίζοντο ἐξ αἰτίας τῆ
Ἄρειας ἢ Μελετίας, μὲ ἰδίαν Ἐπιστολὴν ἐπαρκαλῆσαι τοὺς Χριστιανούς νὰ
ἀποβλέψωσιν εἰς τὸν Θεόν, ἢ νὰ ἀφήσωσιν εἰς αὐτὸν τὸν Θεὸν τῆς κτί-
σεως, ἢ νὰ ἀγαπήσωσιν ἀλλήλους, ἢ ὅσοι τὰς ἐπιβαλεῖνται νὰ τὰς διώ-
κωσι μὲ τὴν ὁμόνοιαν ὅσον δύνανται.

§. 5. Τῷ δὲ Ἀθανασίῳ ὕστερον ἀπὸ πολλὰς ἄλλας **συκοφαντίας** ἢ
πῶ διεσπάρησαν κατ' αὐτῶν, ἢ τότε ἐφανερῶδησαν εἰς τὸν **Ἀυτοκράτορα**
ἀπὸ τῶν Αἰρετικῶν, ἀντέγραψε ἢ νὰ ἐπιμεληθῆται τὴν Ἱερουσαλὴν. νὰ ἐπι-
» ρατῆ, ἢ νὰ φροντίσῃ διὰ τὴν εὐταξίαν, ἢ εὐσέβειαν τῶ λαοῦ, ἢ ὡς
» ἡδὲν νὰ λογίζεται τὰς ἐπιβλάς τῶν Μελετιανῶν ἐπειδὴ ἢ αὐτὸς ὁ
» **Βασιλεὺς** καταλάβε, πῶς ὁ φθόνος τὰς ἐπαρκαλίσεν εἰς τοιαύτας ψευ-
» δεις, ἢ σπασλασμέναις συκοφαντίας, ἢ τὰς δοξάβης, ἢ συγχύσει τῆς
» Ἐκκλησίας νὰ μὴ ἀφίη ὁ **Βασιλεὺς**, νὰ γίνωται, ἀλλὰ νὰ εἶναι δικα-
» ρῆς κατὰ τὰς πολιτικὰς νόμους, ἂν δὲν ἠσυχάσωσι, ἢ νὰ ἐδικαιωθῶ
» ἀπὸ αὐτῶν, ὅπῃ ἔχι μόνον ἀδικίαι ἐπιβαλεῖνται τὰς ἐξώμας ἀλλὰ
» ἢ τὴν εὐταξίαν τῆς Ἐκκλησίας, καὶ εὐσέβειαν λυμάνουσαι ἀνοσίῳι.
Ἐπείραξε δὲ πρὸς τὰτοις, ὅτι αὐτὴ ἡ Ἐπιστολὴ νὰ ἀναγνωθῆ παρήγορη
εἰς τὸν λαόν. Ἀπὸ ἐκείνου τὸν καιρὸν φοβηθέντες οἱ τὰ τῶ Μελετίῳ φρο-
νῶντες διὰ τὰς ἀπειλάς τὰ **Ἀυτοκράτορος** ἠσύχως ἐπολιτεύοντο. Οὗτα
δὲ ἐν εἰρήνῃ πᾶσα ἢ κατ' Αἰγύπτιον Ἐκκλησίᾳ, καὶ ἀπὸ τούτων ἀρχιεπί-
σκοπὸν ἐννομένη, κατ' ἡμέραν ἤξανεν. Ἀνάγνωσι τὸν Σωζόμενον. (α)

ΚΕΦΑ

(α) Ἐκκλησ. Γραμ. βιβλ. Β' κεφ. νῆ κ' ἢ κγ



Συμπέρασμα

- *Η χρήση εργαλείων Γλωσσικής Τεχνολογίας οδήγησε σε σημαντική βελτίωση των αποτελεσμάτων αναζήτησης.*