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 (16ου – 19ου αιω.)
  - Χειρόγραφα (11ου – 19ου αιω.)
- Εργα:
  - Πατριάρχου Φωτίου, Νικοδήμου του Αγιορείτου, Αθανασίου του Παρίου, κ.α.
  - Ομήρου, Θουκυδίδη
  - Γραμματικής, Μαθηματικών, Γεωγραφίας κ.α.

Ψηφιοποίηση της βιβλιοθήκης

- 30ος 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
- Εχουν 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 A.E.
  - 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 (17ου -18ου 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.

Κριτήρια επιλογής λέξεων-κλειδιών

- Αποκλεισμός ιδιαίτερα πλούσιων και πολύτυπων μορφολογικών παραδειγμάτων (κυρίως ρημάτων)
- Συχνότητα εμφάνισης ανά κείμενο, εποχή και κειμενικό είδος
- Βαρύτητα λέξεων από φιλολογικής-ιστορικής-θεολογικής απόψεως
- Ενδιαφέρον από γλωσσολογικής απόψεως,
  - έμφαση σε γραμματικά και ονοματικά στοιχεία τα οποία εμφανίζουν εξέλιξη/μεταβολή κατά τις υπό εξέταση περιόδους και μπορούν να αποτελέσουν αναγνωριστικές ενδείξεις για τον χαρακτηρισμό ενός κειμένου ως δημώδους.

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.)

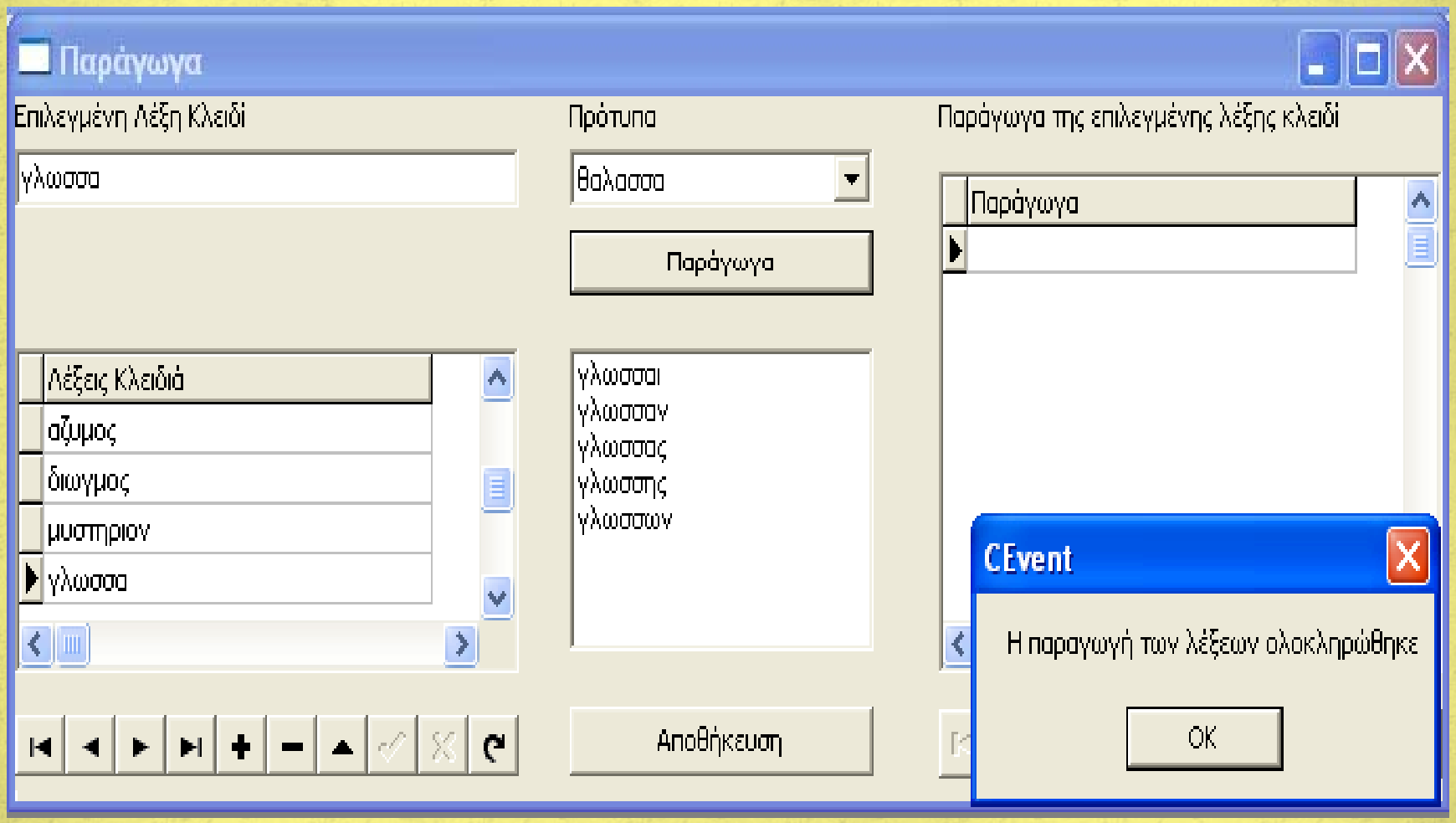
Μορφολογική Γεννήτρια (συν.)

- 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



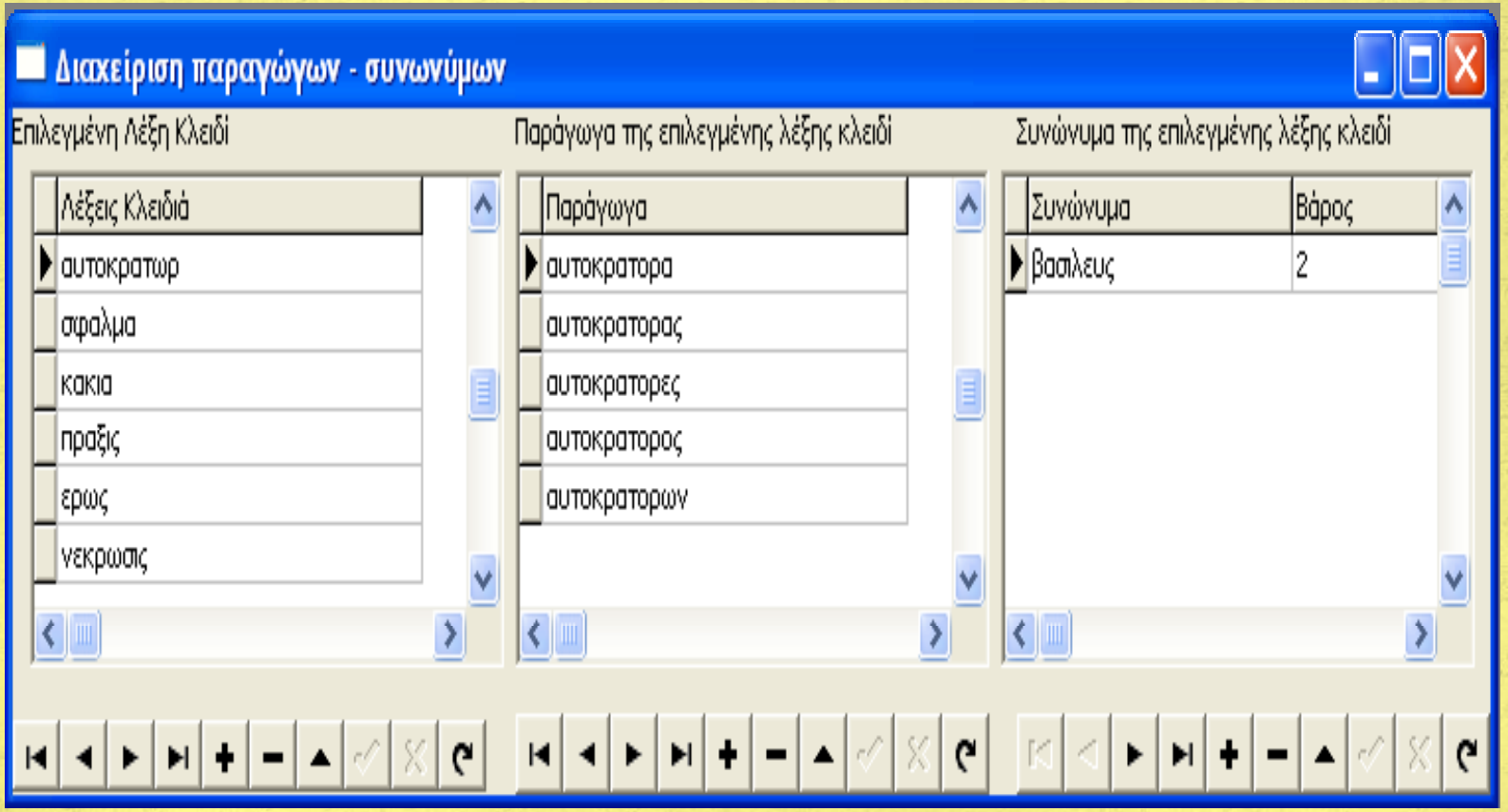
Πρόσβαση στο σημασιολογικό περιεχόμενο

- 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.

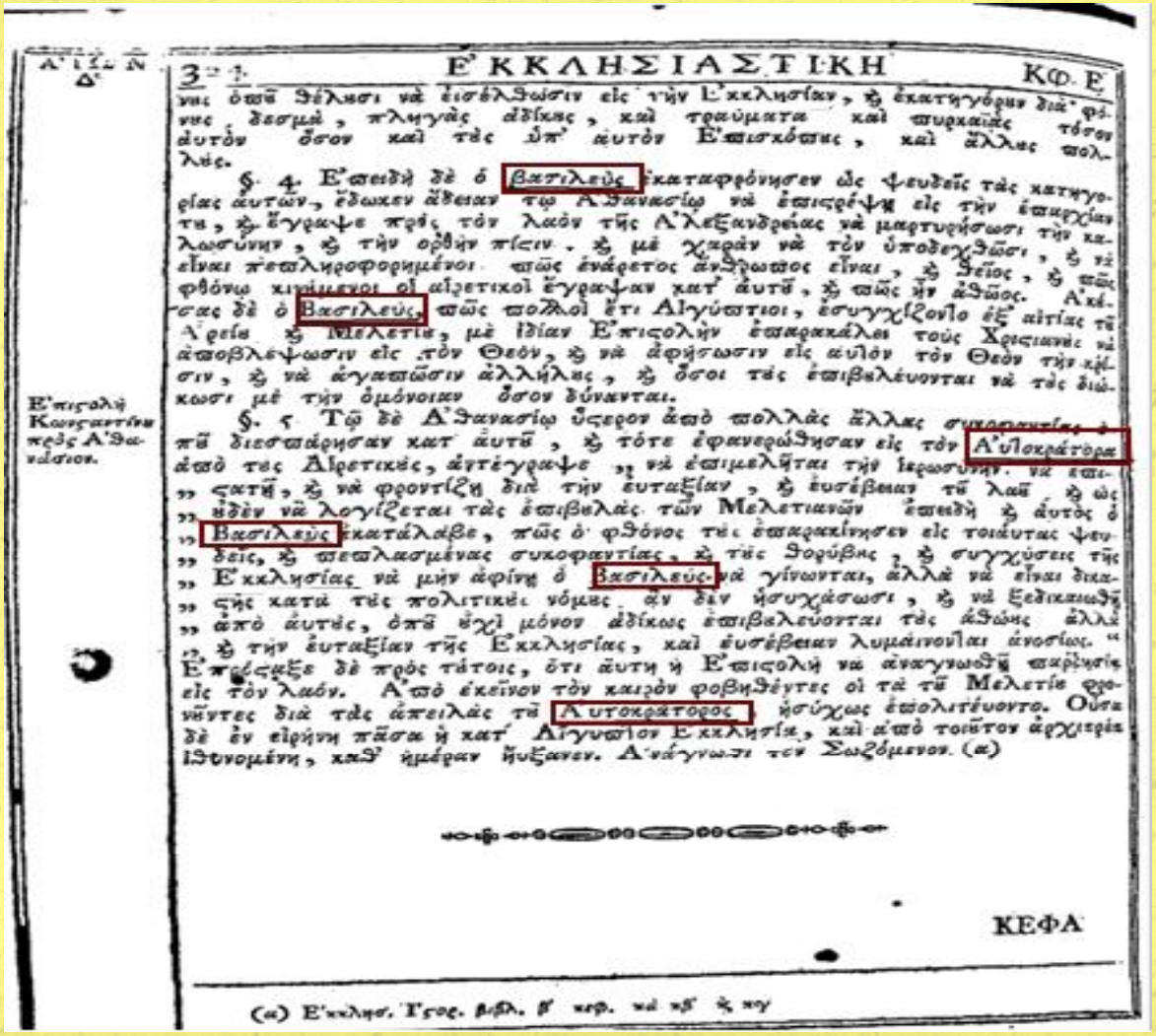
Πρόσβαση στο σημασιολογικό περιεχόμενο (συν.)

- 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).

Παράδειγμα: Εξαγωγή κλιτών μορφών και σταθμισμένων συνωνύμων



Παράδειγμα : Χρήση των κλιτών μορφών και συνωνύμων στη διαδικασία εντοπισμού λέξεων



Α'ΤΙΑΝ Δ'

3 4 Ε'ΚΚΛΗΣΙΑΣΤΙΚΗ ΚΦ. Ε

νας ὁπῶ θέλωσι νὰ εἰσέλθωσιν εἰς τὴν Ἐκκλησίαν, καὶ ἐκατηγόρων διὰ φό-
νους, δεσμὰ, πληγὰς ἀδίκως, καὶ τραύματα καὶ πυρκαϊὰς τόσον
αὐτὸν ὅσον καὶ τὰς ὑπ' αὐτὸν Ἐπισκόπους, καὶ ἄλλας πολ-
λάς.

§. 4 Ἐπειδὴ δὲ ὁ βασιλεὺς ἐκαταφρόνησεν ὡς ψευδεῖς τὰς κατηγο-
ρίας αὐτῶν, ἔδωκεν ἄδειαν τῷ Ἀθανασίῳ νὰ ἐπιστρέψῃ εἰς τὴν ἐπαρχίαν
του, καὶ ἔγραψε πρὸς τὸν λαὸν τῆς Ἀλεξανδρείας νὰ μαρτυρήσωσι τὴν κα-
λωσύνην, καὶ τὴν ὀρθὴν πίστιν. καὶ μὲ χαρὰν νὰ τὸν ὑποδεχθῶσι, καὶ νὰ
εἶναι πεπληροφορημένοι πῶς ἐνάρετος ἄνθρωπος εἶναι, καὶ θεῖος, καὶ πῶς
φθόνῳ κινούμενοι οἱ αἱρετικοὶ ἔγραψαν κατ' αὐτοῦ, καὶ πῶς ἦν ἀθῶος. Ἀκέ-
σας δὲ ὁ βασιλεὺς, πῶς πολλοὶ ἔτι Αἰγύπτιοι, ἐσυγχίζοντο ἐξ αἰτίας τῆς
Ἀρείου καὶ Μελετίου, μὲ ἰδίαν Ἐπιστολὴν ἐπαρακάλει τοὺς Χριστιανὲς νὰ
ἀποβλέψωσιν εἰς τὸν Θεὸν, καὶ νὰ ἀφήσωσιν εἰς αὐτὸν τὸν Θεὸν τὴν κρί-
σιν, καὶ νὰ ἀγαπῶσιν ἀλλήλους, καὶ ὅσοι τὰς ἐπιβουλεύονται νὰ τὰς διώ-
κωσι μὲ τὴν ὁμόνοιαν ὅσον δύνανται.

Ἐπιστολὴ Κωνσταντίνου πρὸς Ἀθανάσιον.

§. 5 Τῷ δὲ Ἀθανασίῳ ὕστερον ἀπὸ πολλὰς ἄλλας συκοφαντίας, καὶ
πῶς διεσπάρησαν κατ' αὐτοῦ, καὶ τότε ἐφανερώθησαν εἰς τὸν Αὐτοκράτορα
ἀπὸ τοὺς Αἱρετικὲς, ἀντέγραψε „ νὰ ἐπιμελῆται τὴν ἱερωσύνην. νὰ ἐπι-
„ στατῇ, καὶ νὰ φροντίζῃ διὰ τὴν εὐταξίαν, καὶ εὐσέβειαν τοῦ λαοῦ καὶ ὡς
„ ἠδὲν νὰ λογίζεται τὰς ἐπιβουλὰς τῶν Μελετιανῶν ἐπειδὴ καὶ αὐτὸς ὁ
„ βασιλεὺς ἐκατάλαβε, πῶς ὁ φθόνος τὸν ἐπαρακίνησεν εἰς τοιούτους ψευ-
„ δεῖς, καὶ πεπλασμένας συκοφαντίας, καὶ τὰς θορύβους, καὶ συγχύσεις τῆς
„ Ἐκκλησίας νὰ μὴν ἀφίνῃ ὁ βασιλεὺς νὰ γίνωνται, ἀλλὰ νὰ εἶναι διακ-
„ ρῆς κατὰ τὰς πολιτικὲς νόμους ἂν δὲν ἡσυχάσωσι, καὶ νὰ ξεδικαιωθῇ
„ ἀπὸ αὐτὸς, ὁπῶ οὐχὶ μόνον ἀδίκως ἐπιβουλεύονται τὰς ἀθῶς ἀλλὰ
„ καὶ τὴν εὐταξίαν τῆς Ἐκκλησίας, καὶ εὐσέβειαν λυμαίνονται ἀνοσίως. "
Ἐπρόσθεσε δὲ πρὸς τούτοις, ὅτι αὐτὴ ἡ Ἐπιστολὴ νὰ ἀναγνωσθῇ παρρησίᾳ
εἰς τὸν λαόν. Ἀπὸ ἐκεῖνον τὸν καιρὸν φοβηθέντες οἱ τὰ τοῦ Μελετίου φρο-
νοῦντες διὰ τὰς ἀπειλὰς τοῦ Αὐτοκράτορος ἡσύχως ἐπολιτεύοντο. Οὕτω
δὲ ἐν εἰρήνῃ πᾶσα ἡ κατ' Αἴγυπτον Ἐκκλησία, καὶ ἀπὸ τοιοῦτον ἀρχιερέα
ἰθυνομένη, καθ' ἡμέραν ἠυξανεν. Ἀνάγνωσι τὸν Σωζόμενον. (α)

ΚΕΦΑ

(α) Ἐκκλησ. Γρος. βιβλ. β' κεφ. κα' κβ' & κγ

Συμπέρασμα

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