ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών

Εισαγωγή στην Ανάλυση Γλωσσικών Δεδομένων

Ενότητα 1: Η ελληνική γλώσσα μέσα από αριθμούς: Μετρήσεις και στατιστική στην υπηρεσία της γλωσσολογίας

Γεώργιος Κ. Μικρός
Φιλοσοφική Σχολή
Τμήμα Ιταλικής Γλώσσας και Φιλολογίας

Ποσοτική γλωσσολογία

- Ποσοτική Γλωσσολογία (ΠΓ) είναι ο κλάδος εκείνος της Γλωσσολογίας που ασχολείται με την ποσοτική ανάλυση της γλωσσικής δομής και τη γλωσσολογική ερμηνεία της.
- Η ποσοτική ανάλυση χρησιμοποιείται για να ολοκληρωθεί η ποιοτική ανάλυση που διεξάγει ο κλάδος της θεωρητικής γλωσσολογίας. Γενικότερα, η χρήση ποσοτικών μεθόδων όπως αυτές που θα περιγραφούν παρακάτω λειτουργούν συμπληρωματικά με τις ποιοτικές θεωρήσεις ως προς την κατανόηση του γλωσσικού φαινομένου.

Ηλεκτρονικά Σώματα Κειμένων

- Η ποσοτική αντιμετώπιση της γλωσσικής χρήσης θα ήταν αδύνατη εάν δεν υπήρχαν τα Ηλεκτρονικά Σώματα Κειμένων
- Ορισμοί
  - είναι η συλλογή τμημάτων γλώσσας τα οποία επιλέγονται και διατάσσονται σύμφωνα με συγκεκριμένα γλωσσολογικά κριτήρια έτσι ώστε να χρησιμοποιηθούν ως αντιπροσωπευτικό δείγμα μιας συγκεκριμένης γλώσσας (EAGLES 1996).
  - είναι μια συλλογή κειμένων η οποία είναι κωδικοποιημένη για τυποποιημένες (standardized) και ομοιογενείς εργασίες ανάκτησης γλωσσικής πληροφορίας.

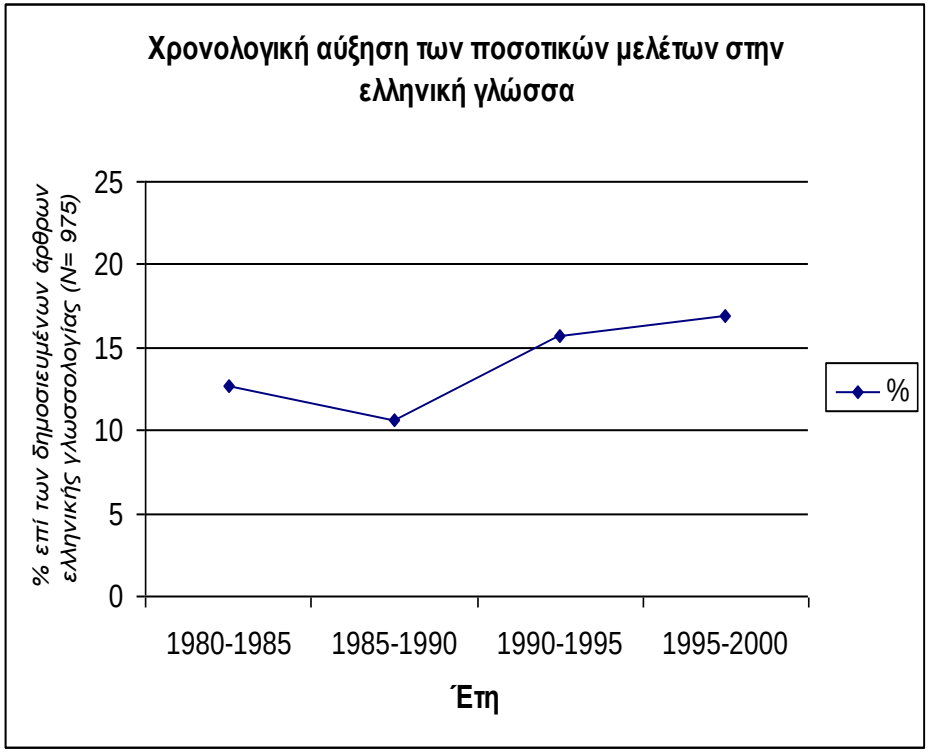
Χαρακτηριστικά της έρευνας που βασίζεται στη χρήση ΗΣΚ

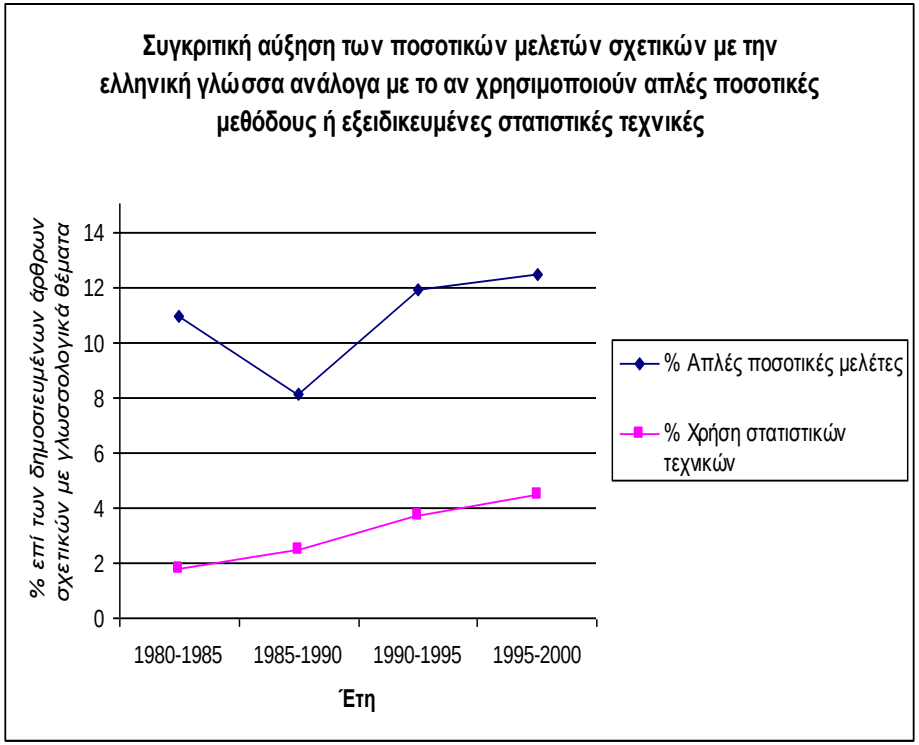
- Άμεση εξάρτηση από τα κείμενα που περιέχονται στο ΗΣΚ
  - Η ποσοτική (μέγεθος) και η ποιοτική (κειμενική ποικιλία) σύσταση του ΗΣΚ διαμορφώνει τα αποτελέσματα που παίρνουμε από αυτό
- Αξιοποίηση των Η/Υ στην επεξεργασία του γλωσσικού υλικού
  - Ταχύτητα και αξιοπιστία κατά την εκτέλεση τυποποιημένων εργασιών γλωσσικής ανάλυσης
- Ποσοτική και ποιοτική προσέγγιση της γλωσσικής χρήσης
  - Η γλωσσική χρήση αντιμετωπίζεται ολιστικά και παρέχονται πληροφορίες τόσο για την ποσοτική δομή, όσο και για την λειτουργική αλληλεπίδραση των γλωσσικών στοιχείων

Ο ρόλος των Η/Υ στην ανάλυση των ΗΣΚ

- Δυνατότητα τεράστιας αποθήκευσης κειμενικών δεδομένων
- Μεγάλη ταχύτητα επεξεργασίας γλωσσικών δεδομένων (Wordsmith: 15.000 λέξεις το δλπτο)
- Συνεπής και «αλάνθαστη» απόδοση σε επαναληπτικές διαδικασίες

Ποσοτικές μελέτες στην ελληνική γλώσσα





Ποσοτική ανάλυση της Νέας Ελληνικής

- Ανάλυση της γλωσσικής χρήσης
  - Λίστες συχνότητας λέξεων
  - Λεξικά σύμπλοκα
  - Γλωσσική ποικιλία
  - Στατιστικά χαρακτηριστικά της γλώσσας
- Υφομετρική ανάλυση κειμένων

Ανάλυση Γλωσσικής Χρήσης: Λίστες Συχνότητας Λεξιλογίου

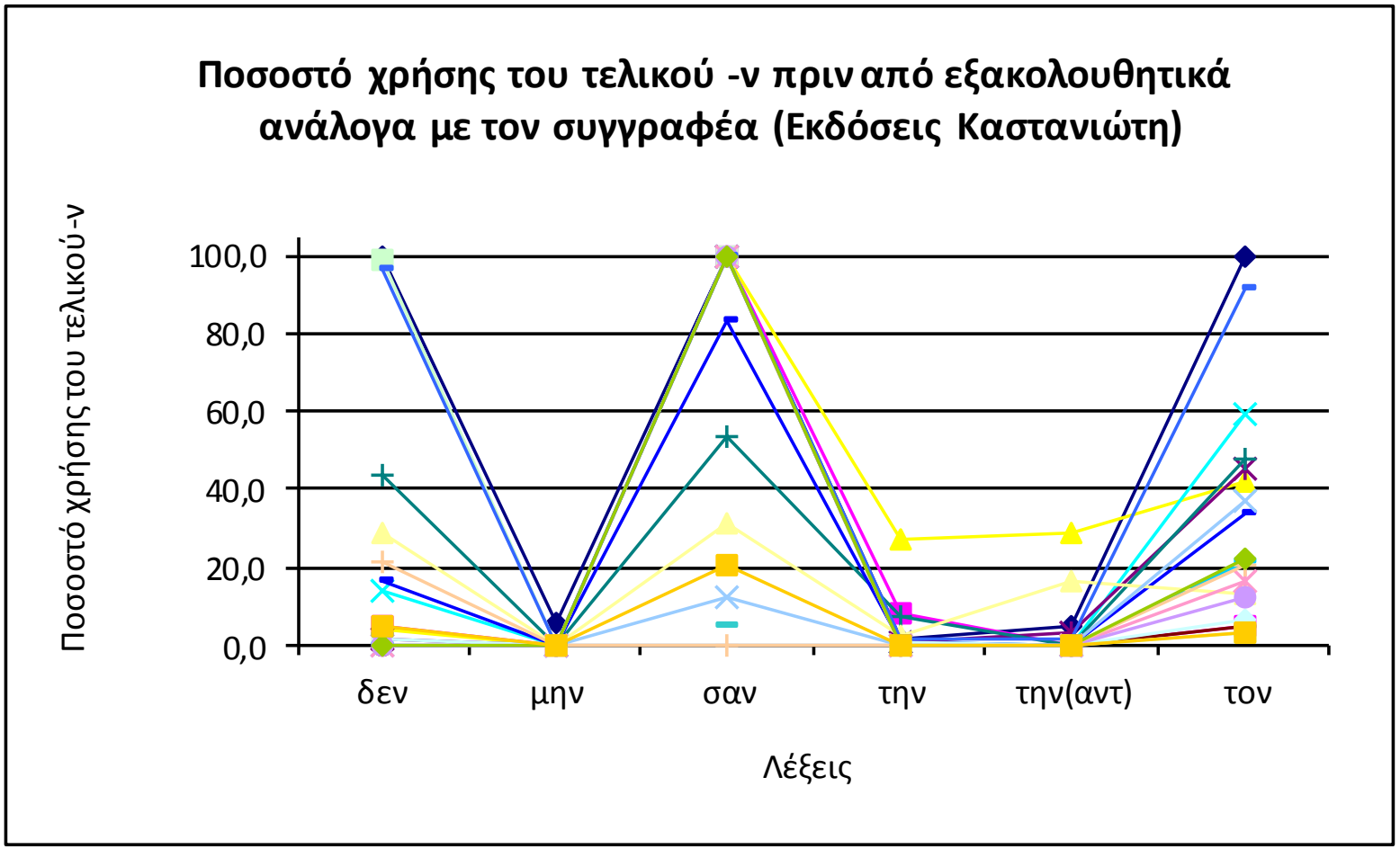
N	Word	Freq.	%
1	και	52.675	2,99
2	το	37.936	2,15
3	του	36.708	2,08
4	να	34.539	1,96
5	της	29.760	1,69
6	η	27.909	1,58
7	που	25.922	1,47
8	την	24.985	1,42
9	με	21.951	1,25
10	από	20.384	1,16
11	για	18.205	1,03
12	τα	18.145	1,03
13	ο	17.423	0,99
14	των	17.366	0,99
15	είναι	15.038	0,85
16	δεν	13.869	0,79
17	τη	12.922	0,73
18	σε	12.819	0,73
19	οι	12.792	0,73
20	τον	12.278	0,70
21	στο	12.223	0,69
22	θα	11.805	0,67
23	τους	11.138	0,63
24	στην	10.573	0,60
25	τις	9.245	0,52
26	ότι	8.629	0,49

N	Word	Freq.
3545	αα	8
3546	ααα	3
3547	άαα	1
3548	άαααα	1
3549	άαααχ	1
3550	αβ	2
3551	αβαθείς	1
3552	αβαθή	2
3553	άβακα	2
3554	αβακούμ	1
3555	αβάνα	1
3556	αβαντάζ	3
3557	αβαπτιστα	2
3558	αβαρίες	1
3559	αβασάνιστα	4
3560	αβασάνιστη	1
3561	αβασάνιστης	1
3562	αβασάνιστο	1
3563	αβάσιμη	1
3564	αβασκαμένους	1
3565	αβάσκαντη	1
3566	αβάστακτη	2
3567	αβάσταχτες	1
3568	αβασταχτη	1
3569	αβάσταχτη	4
3570	αβάσταχτο	5

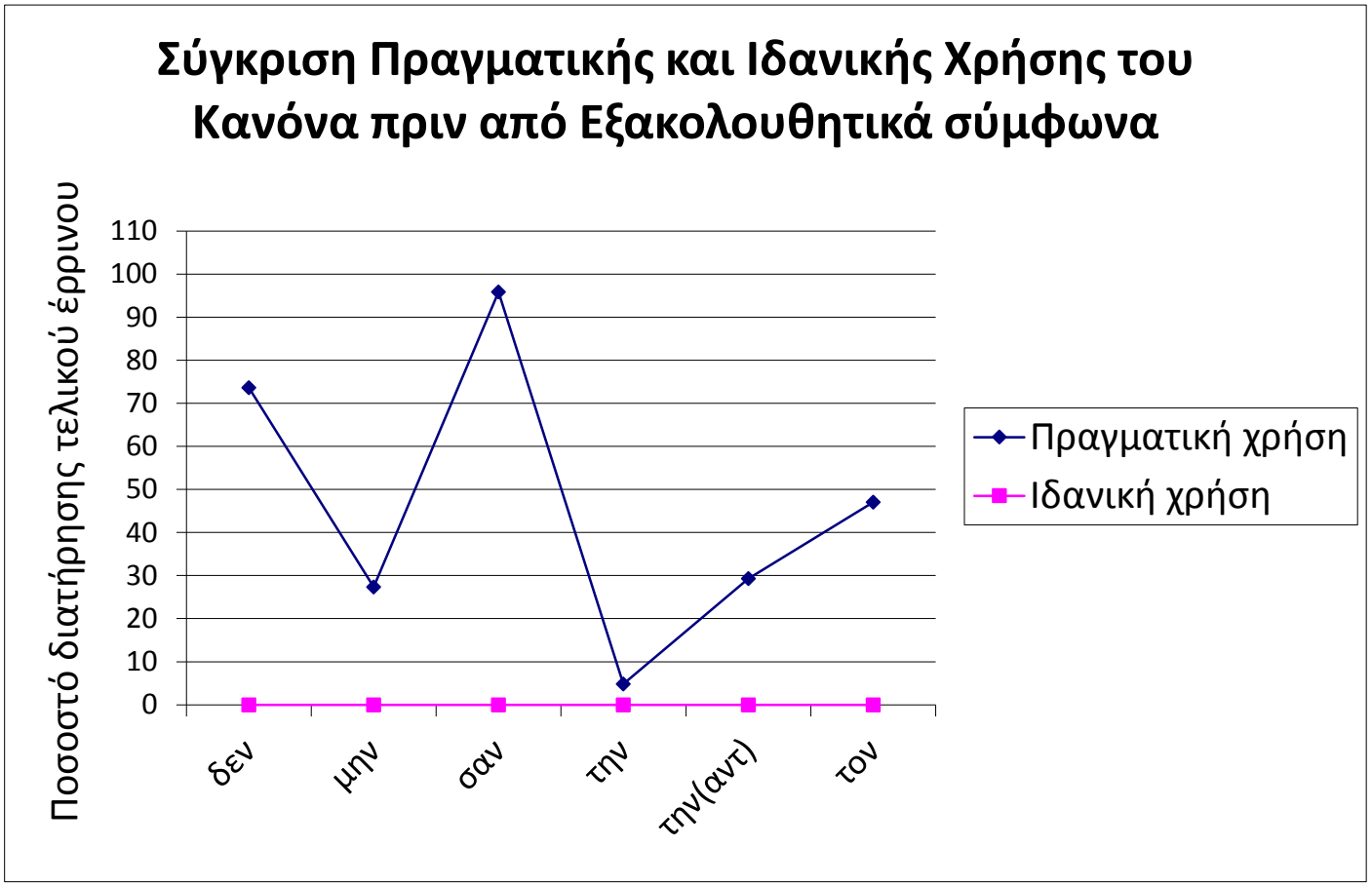
Ανάλυση Γλωσσικής Χρήσης: Τα συχνότερα κύρια ονόματα

Α/Α	Λέξη	Εμφανίσεις	Συχνότητα (τοις χιλίοις)
1	Ελλάδα	30032	0,8754 %
2	Αθήνα	13020	0,3795 %
3	Τουρκία	11944	0,3481 %
4	Ευρώπη	8976	0,2616 %
5	Παπανδρέου	8855	0,2581 %
6	Έλληνας	7324	0,2135 %
7	Σημίτης	6226	0,1815 %
8	Θεσσαλονίκη	5888	0,1716 %
9	Ρέππας	4946	0,1442 %
10	Γιώργος	4837	0,1410 %

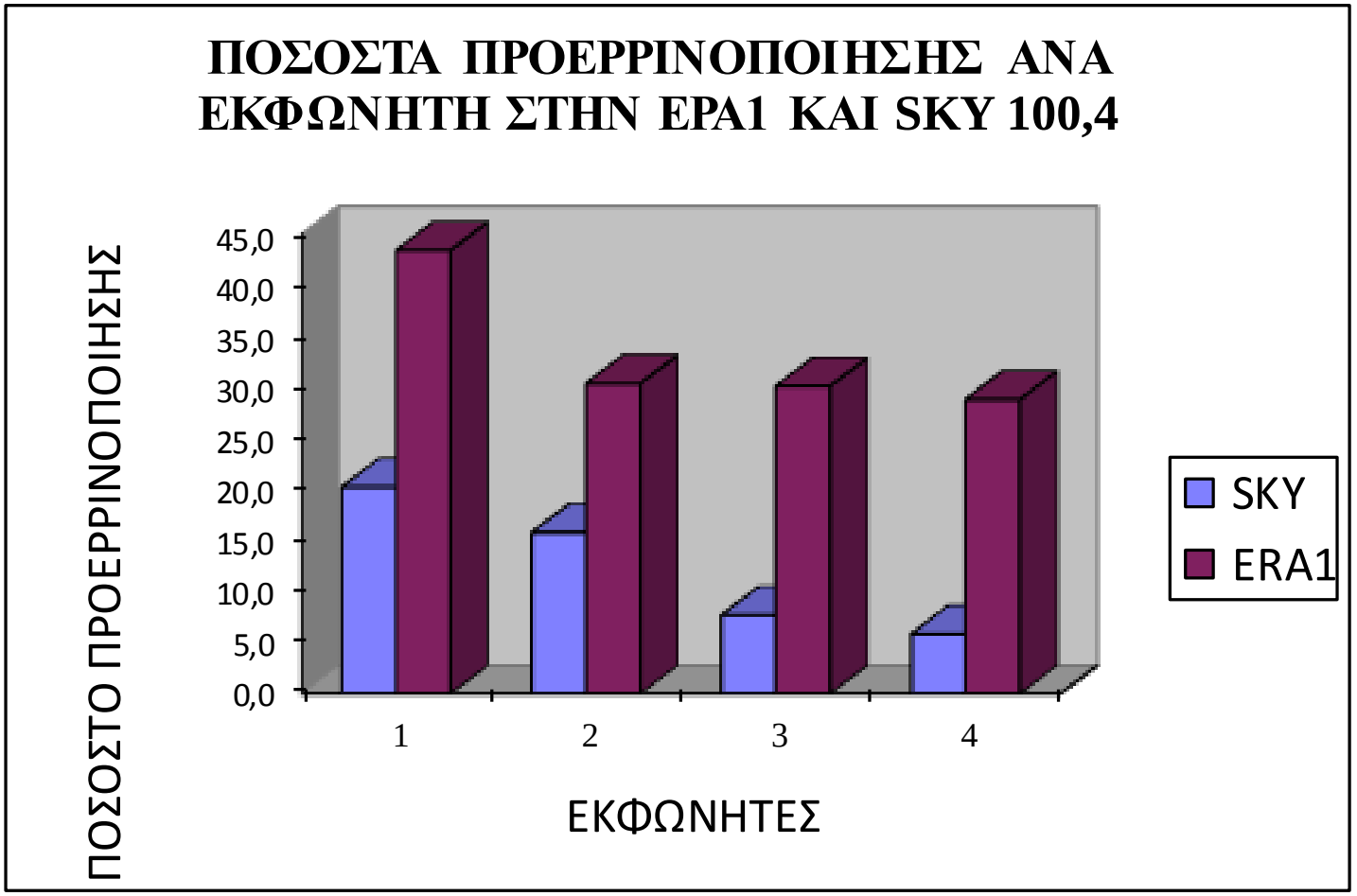
Ανάλυση Γλωσσικής Χρήσης: Γλωσσική Ποικιλία Ι



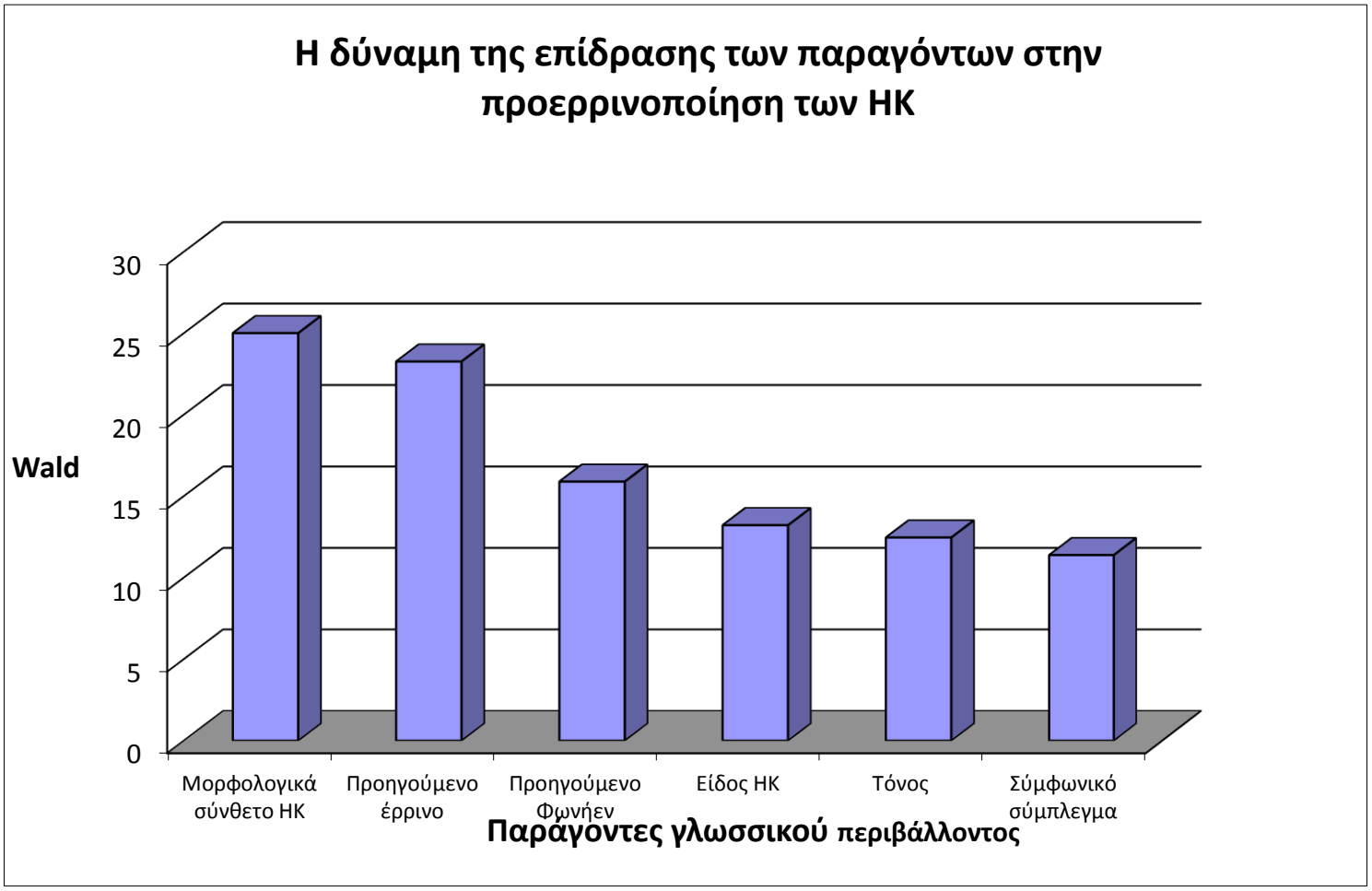
Ανάλυση Γλωσσικής Χρήσης: Γλωσσική Ποικιλία ΙΙ



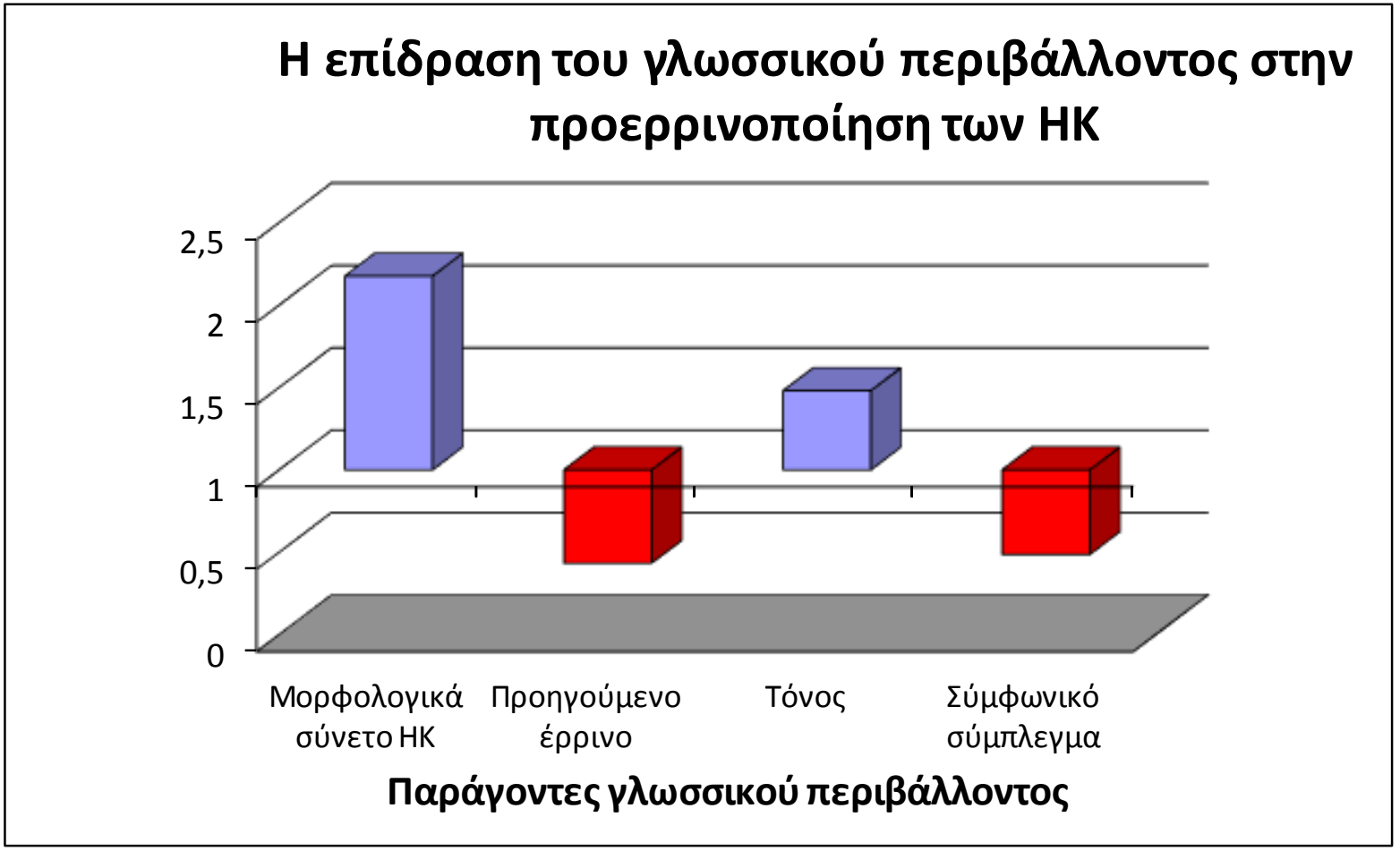
Ανάλυση Γλωσσικής Χρήσης: Γλωσσική Ποικιλία ΙΙΙ



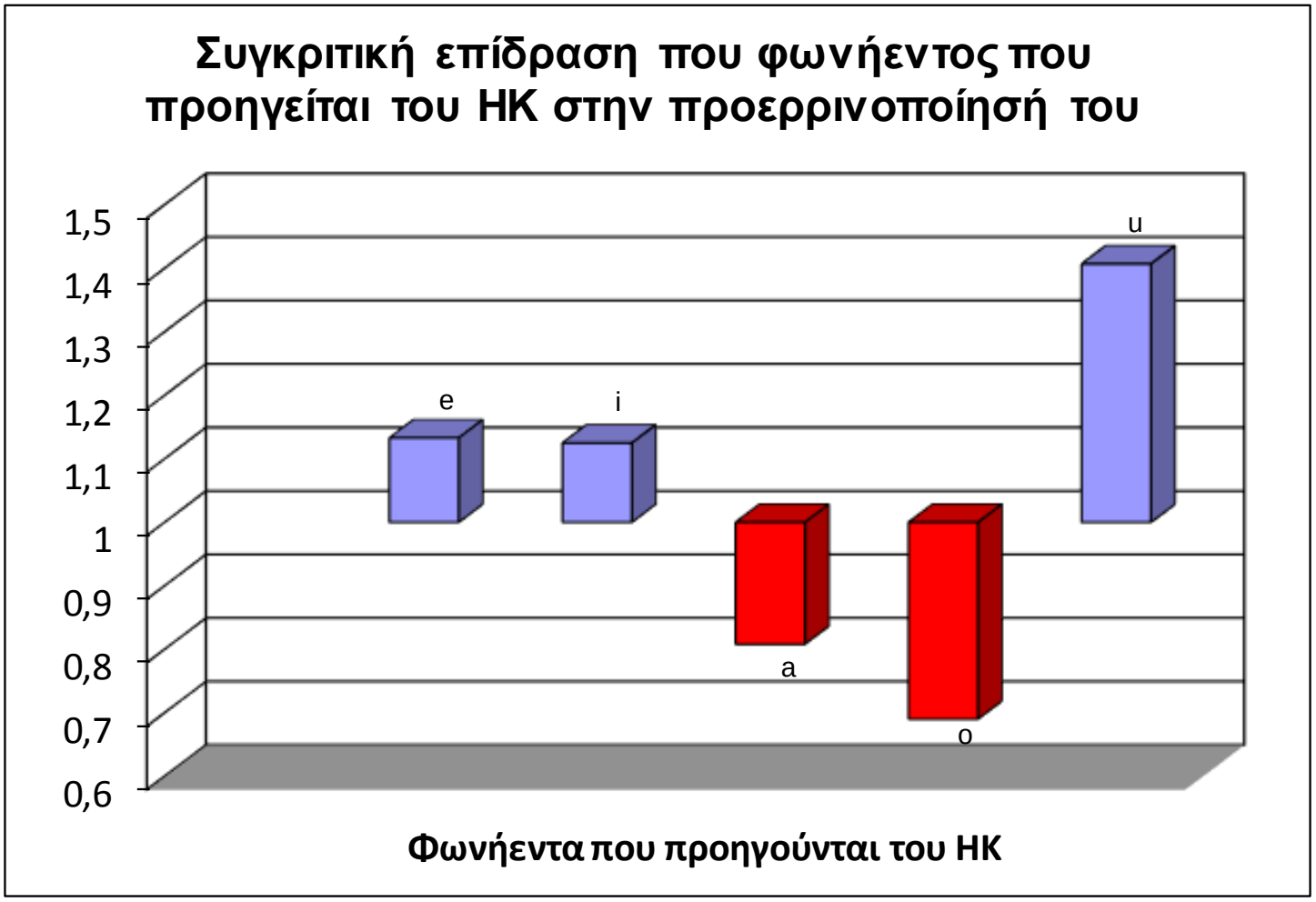
Οι γλωσσικοί παράγοντες που σχετίζονται με την προερρινοποίηση



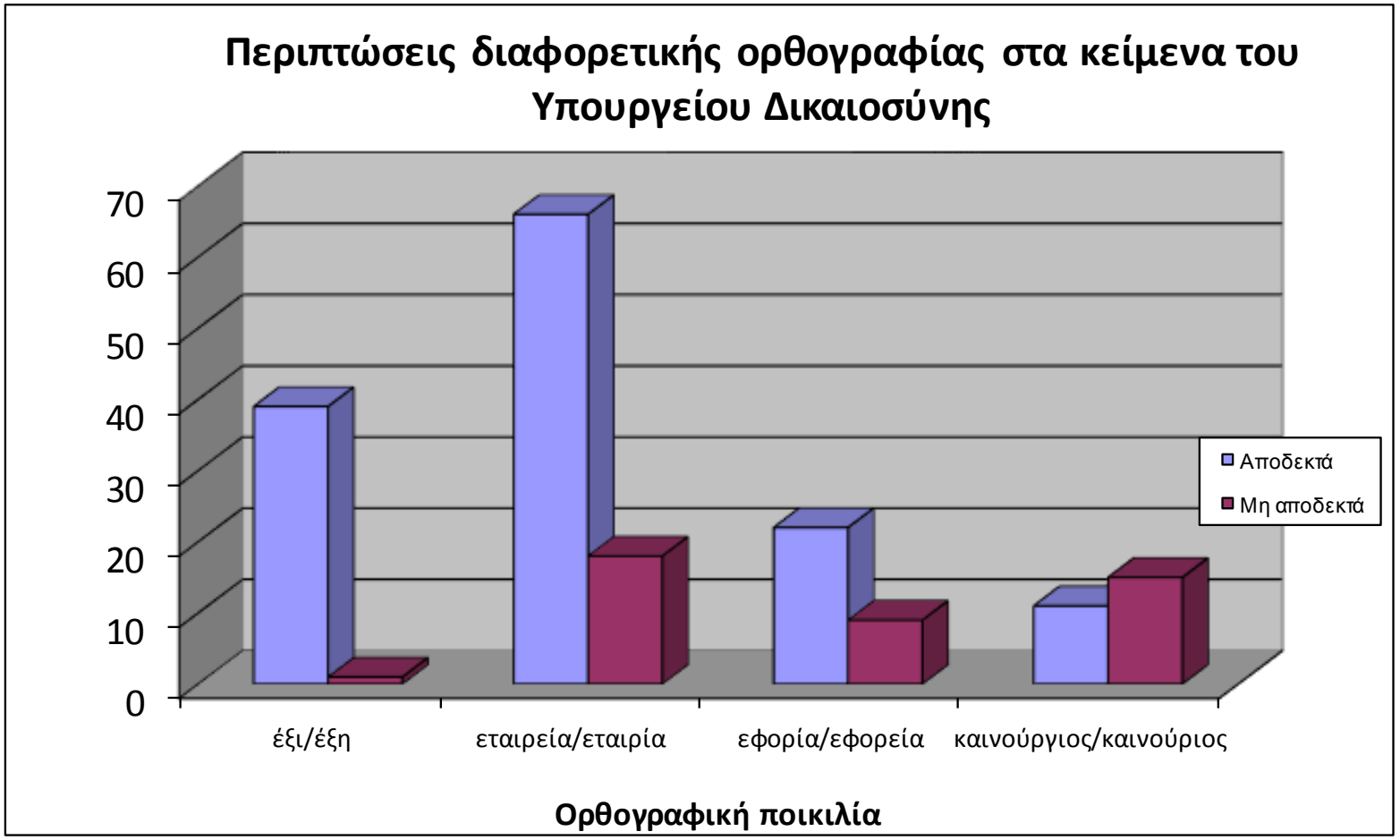
Η επίδραση του γλωσσικού περιβάλλοντος στην προερρινοποίηση των ΗΚ



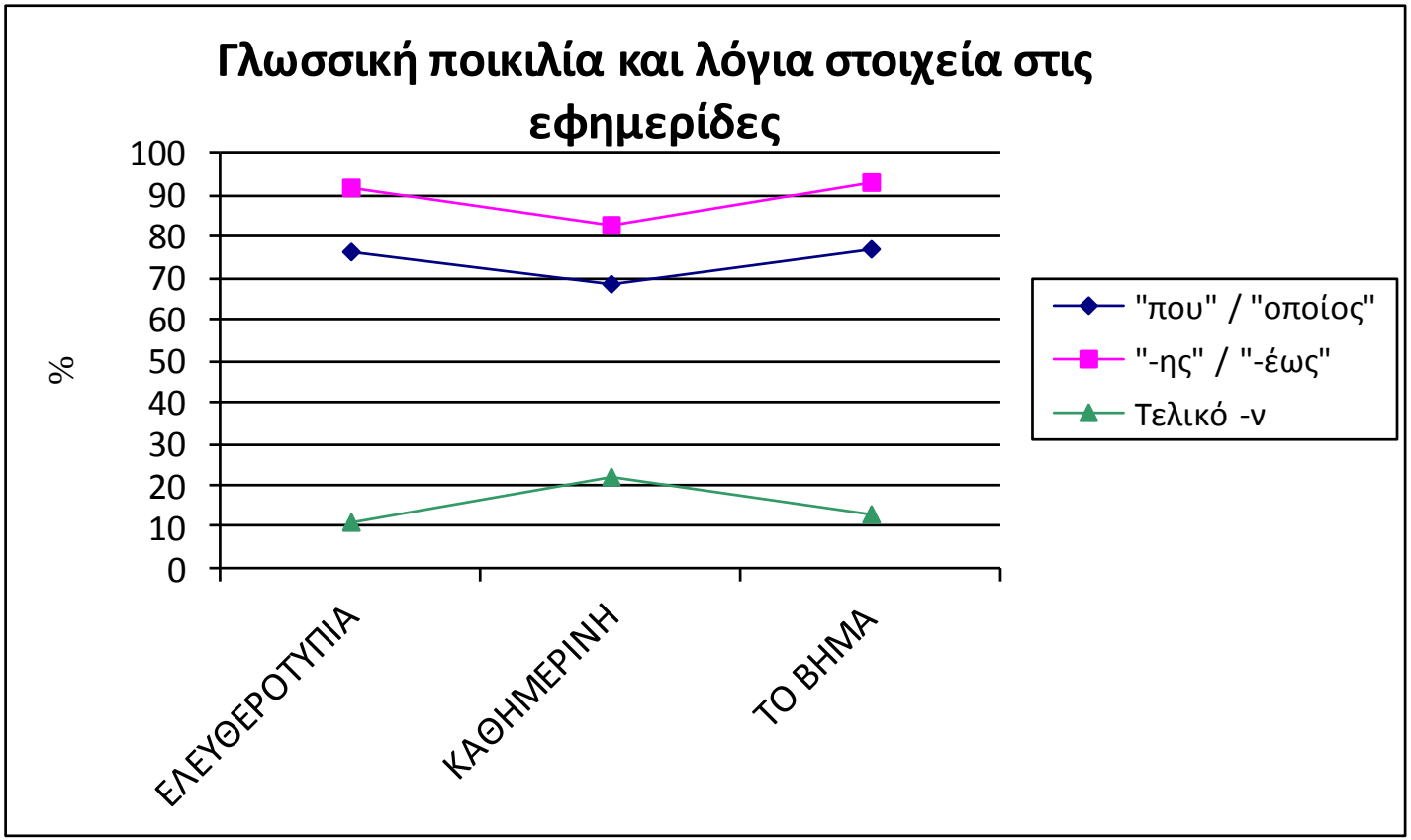
Διερεύνηση της επίδρασης που ασκεί το φωνήεν που προηγείται ενός ΗΚ στην προερρινοποίηση



Ανάλυση Γλωσσικής Χρήσης: Γλωσσική Ποικιλία IV



Ανάλυση Γλωσσικής Χρήσης: Γλωσσική Ποικιλία V



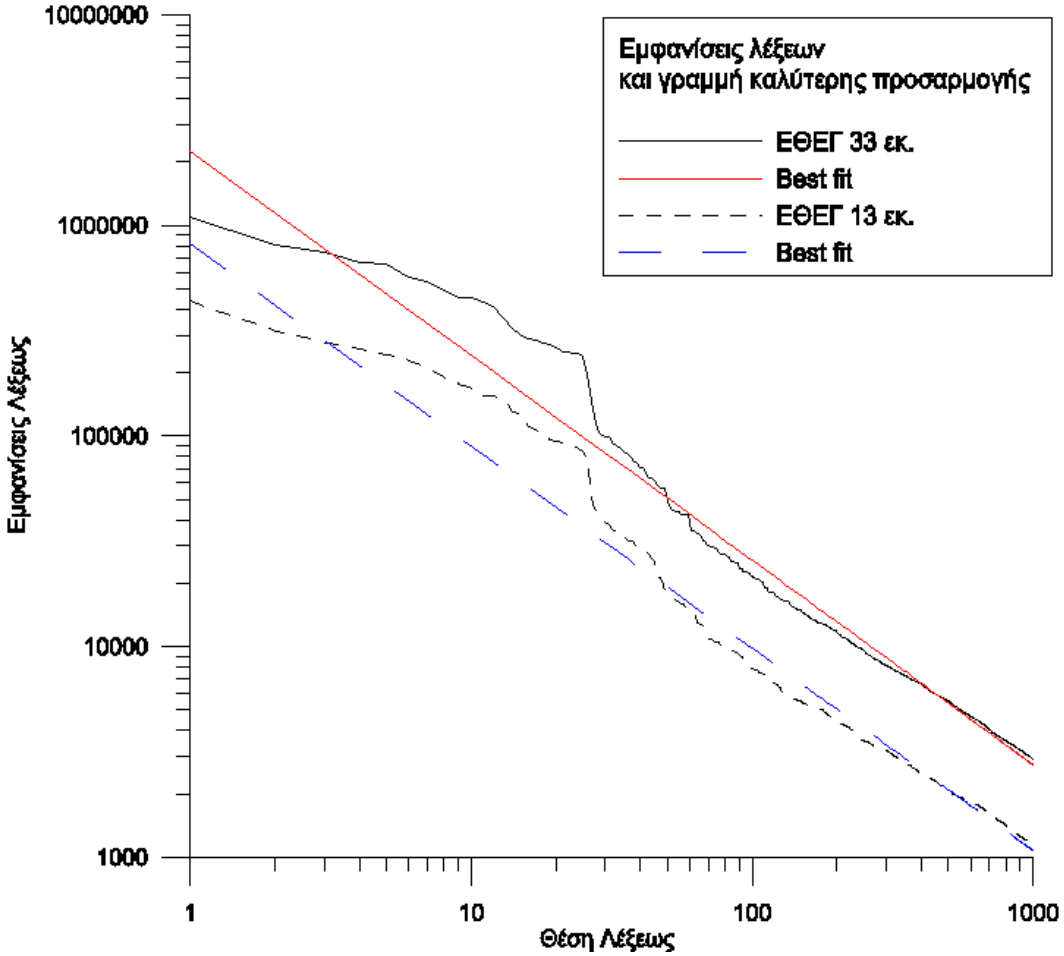
Στατιστικά χαρακτηριστικά της γλώσσας Ι

- Η συχνότητα των λέξεων παρουσιάζει μια «παράξενη» κανονικότητα όταν εξετάζεται σε μια Λίστα Συχνότητας Λεξιλογίου – ΛΣΛ.

Και	1000	Χ	1 = 1000
Του	500	Χ	2 = 1000
Της	333	Χ	3 ≈ 1000

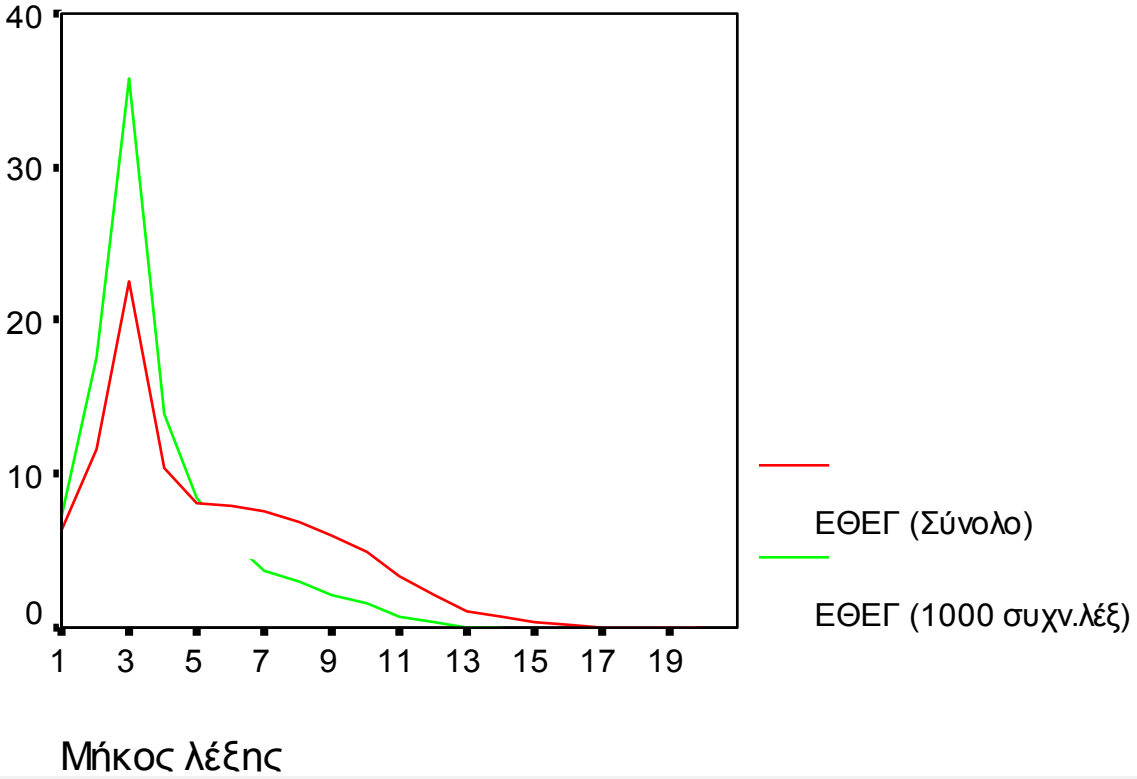
- Το σταθερό γινόμενο (συχνότητα λέξης Χ σειρά κατάταξης σε ΛΣΛ) ονομάστηκε σταθερά του Zipf (c) και αντιπροσωπεύει την μαθηματική αποτύπωση της αρχής της ελάχιστης προσπάθειας που διέπει την ανθρώπινη επικοινωνία.

Στατιστικά χαρακτηριστικά της γλώσσας ΙΙ

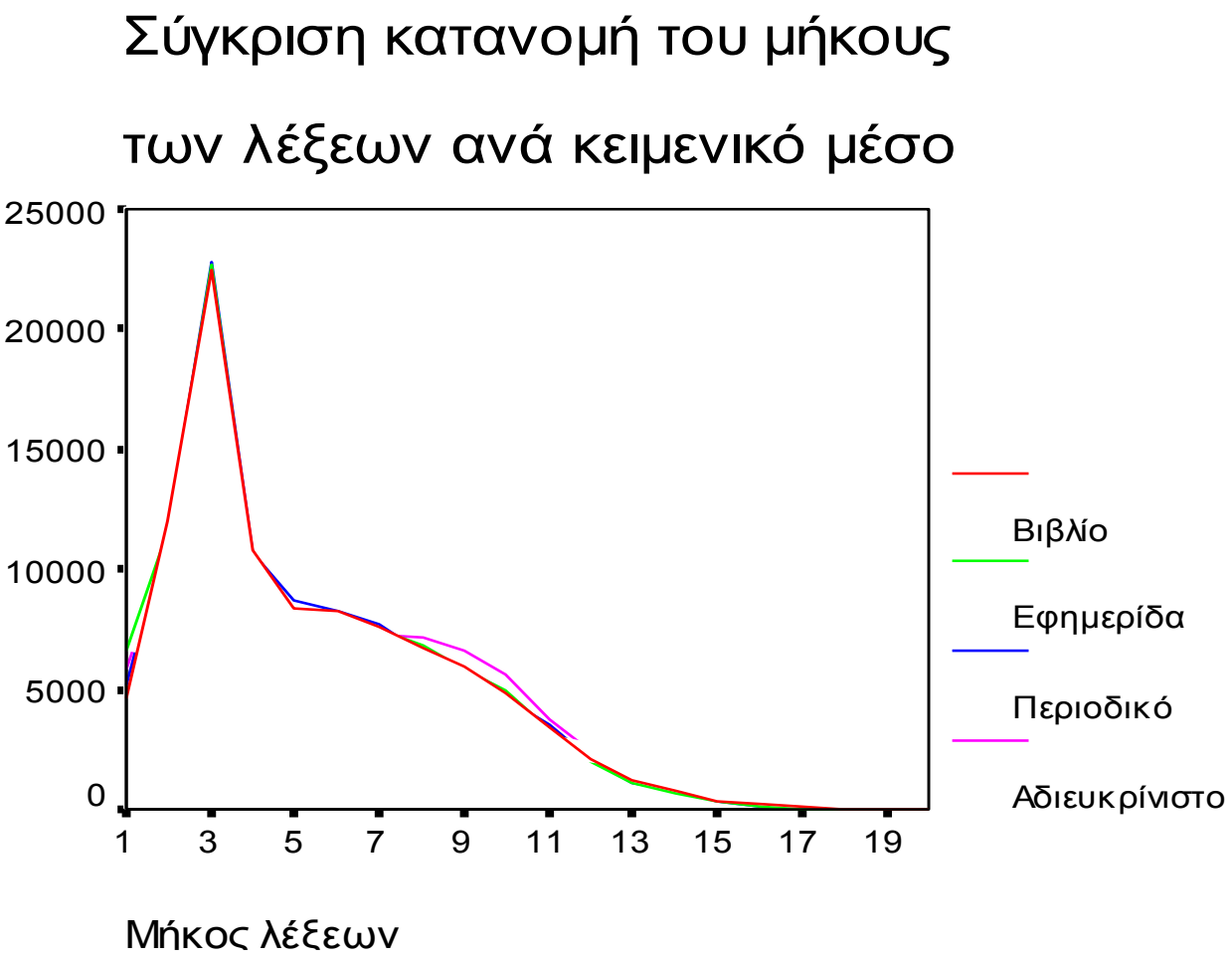


Στατιστικά χαρακτηριστικά της γλώσσας ΙΙΙ

Σύγκριση του μήκους λέξης στις 1000 συχν. λέξεις και στο σύνολο του ΕΘΕΓ



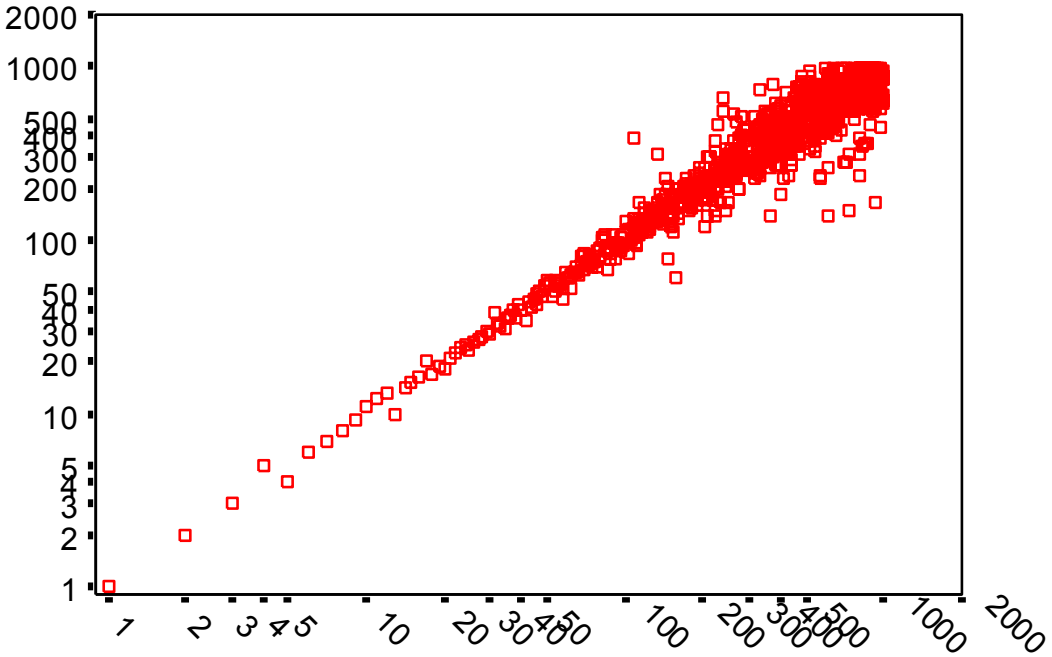
Στατιστικά χαρακτηριστικά της γλώσσας ΙΙΙ



Στατιστικά χαρακτηριστικά της γλώσσας IV

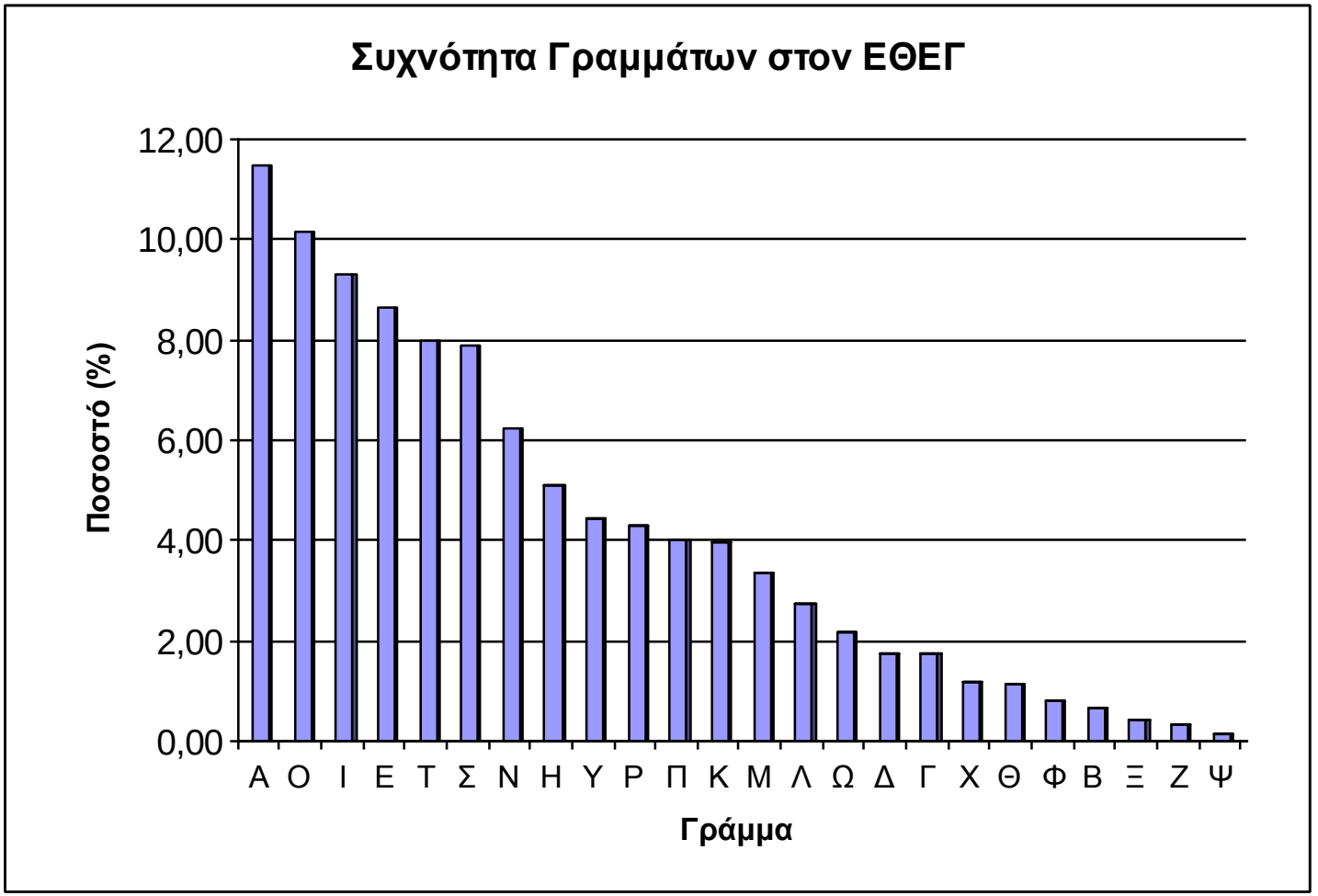
Σχετική θέση των 1000 συχ. λέξεων

στις 2 εκδόσεις του ΕΘΕΓ

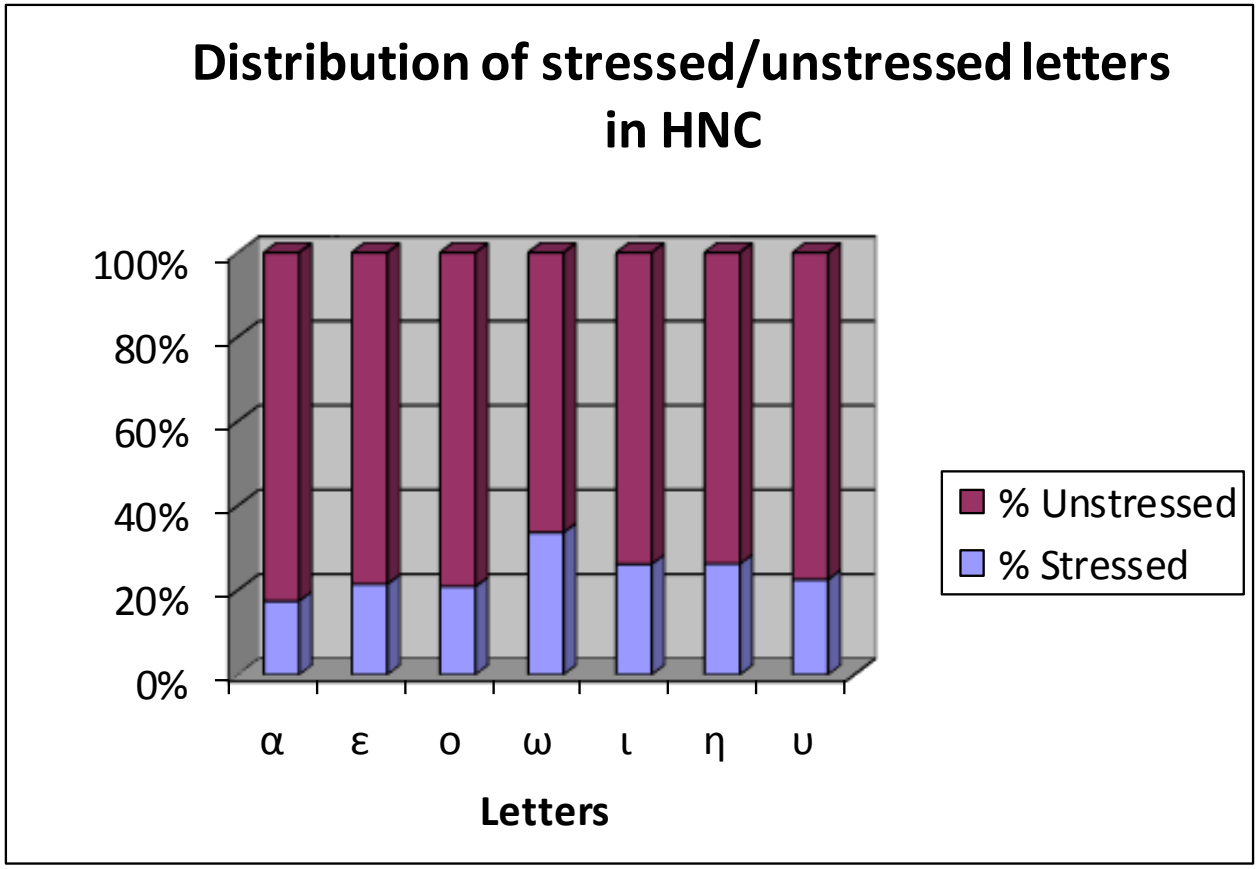


ΕΘΕΓ 13 εκ. λέξεις

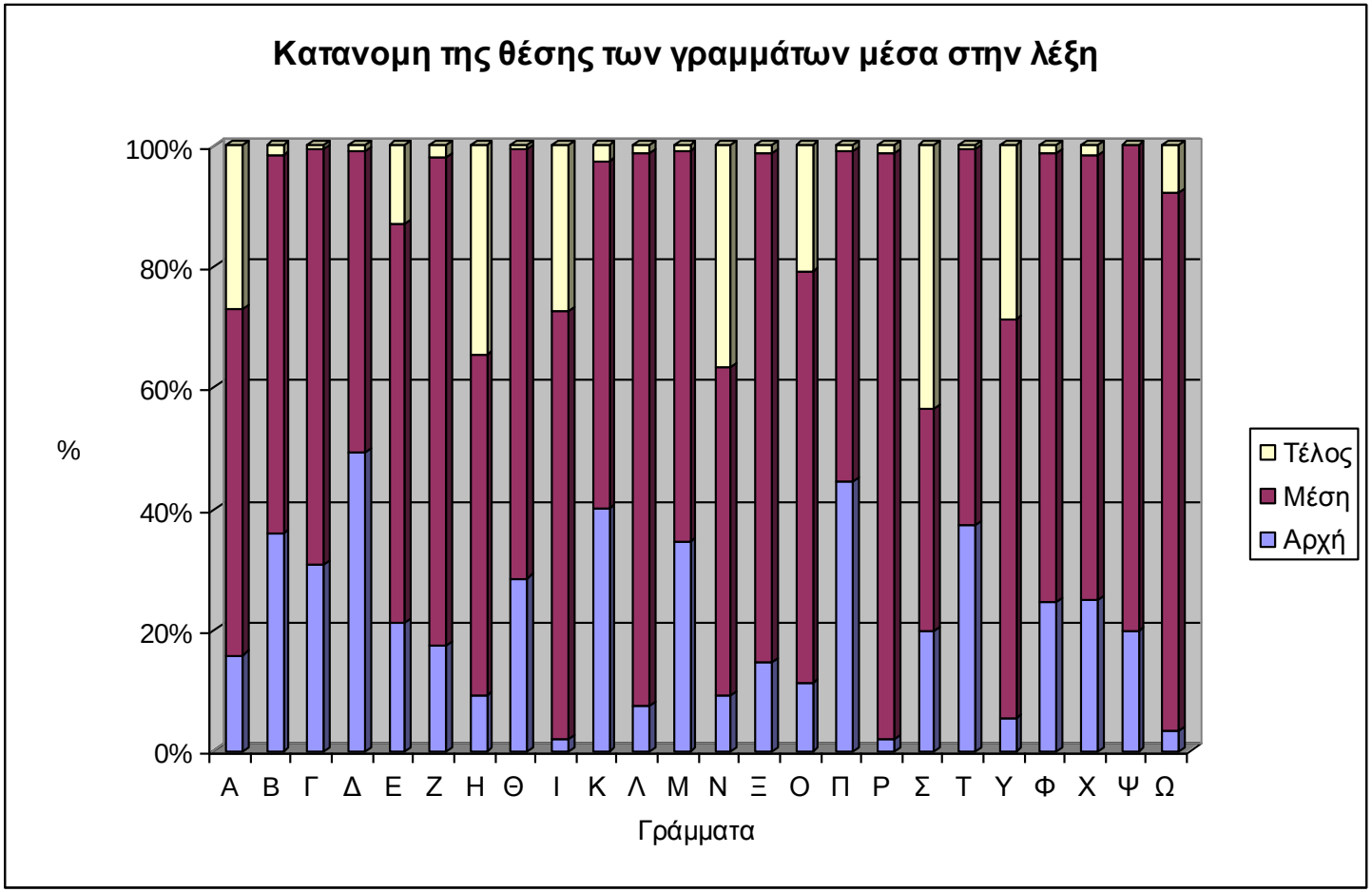
Στατιστικά χαρακτηριστικά της γλώσσας V: Συχνότητες Γραμμάτων



Στατιστικά χαρακτηριστικά της γλώσσας V: Συχνότητες Γραμμάτων



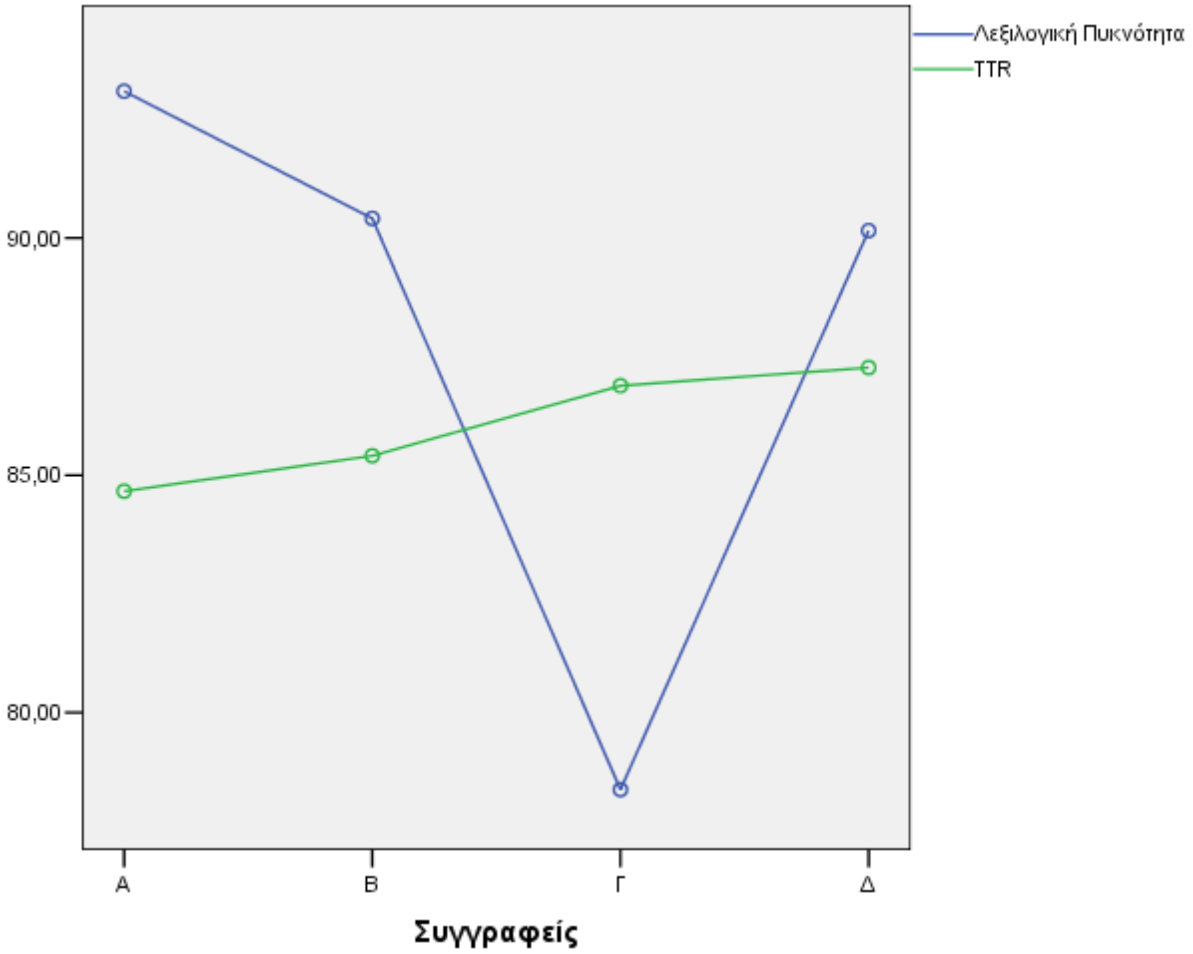
Στατιστικά χαρακτηριστικά της γλώσσας V: Συχνότητες Γραμμάτων



Μερικά βασικά συμπεράσματα

- Οι 1000 συχνότερες λέξεις μένουν σταθερές ανεξαρτήτως του μεγέθους του ΗΣΚ που τις εξετάζουμε (89,5% ομοιότητα)
- Ο νόμος του Zipf ισχύει για τις 1000 συχνότερες λέξεις και λήμματα και εξηγεί την σταθερότητά τους.
- Η κατανομή του μήκους της λέξης ακολουθεί την Negative Binomial.
- Τα πιο συχνά φωνήεντα είναι τα «Α, Ο, Ι» και τα ποιο συχνά σύμφωνα είναι τα «Τ, Σ, Ν».
- 22% των φωνηέντων είναι τονισμένα.

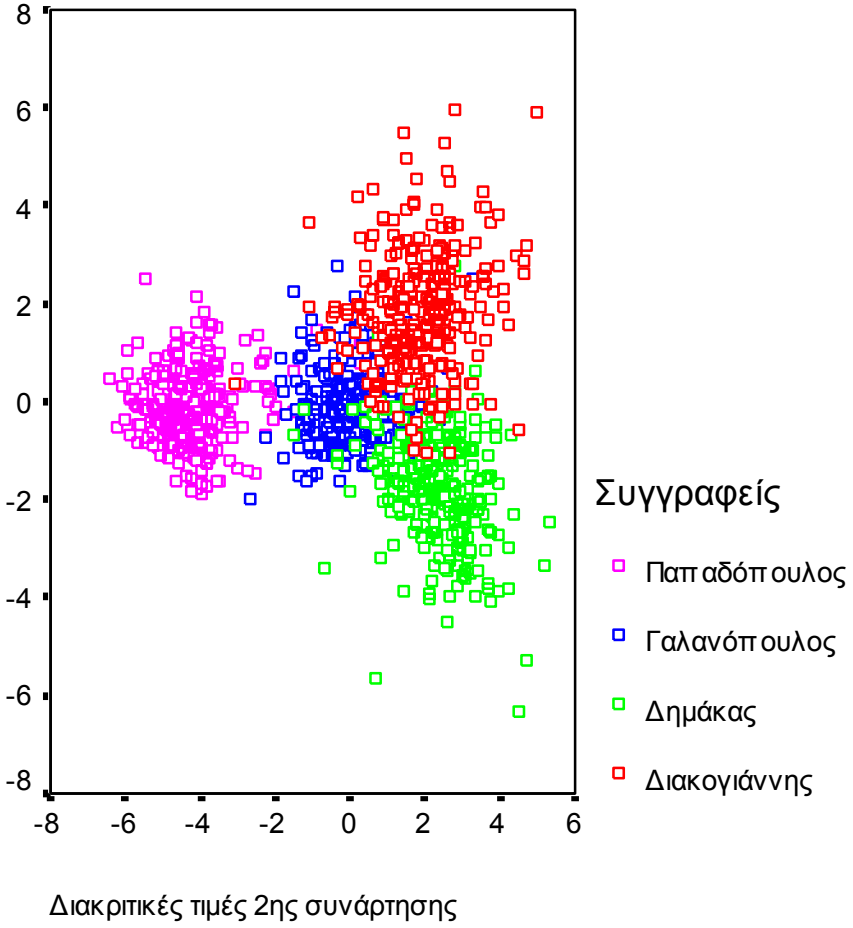
Υφομετρικά χαρακτηριστικά και αναγνώριση αγνώστου συγγραφέα



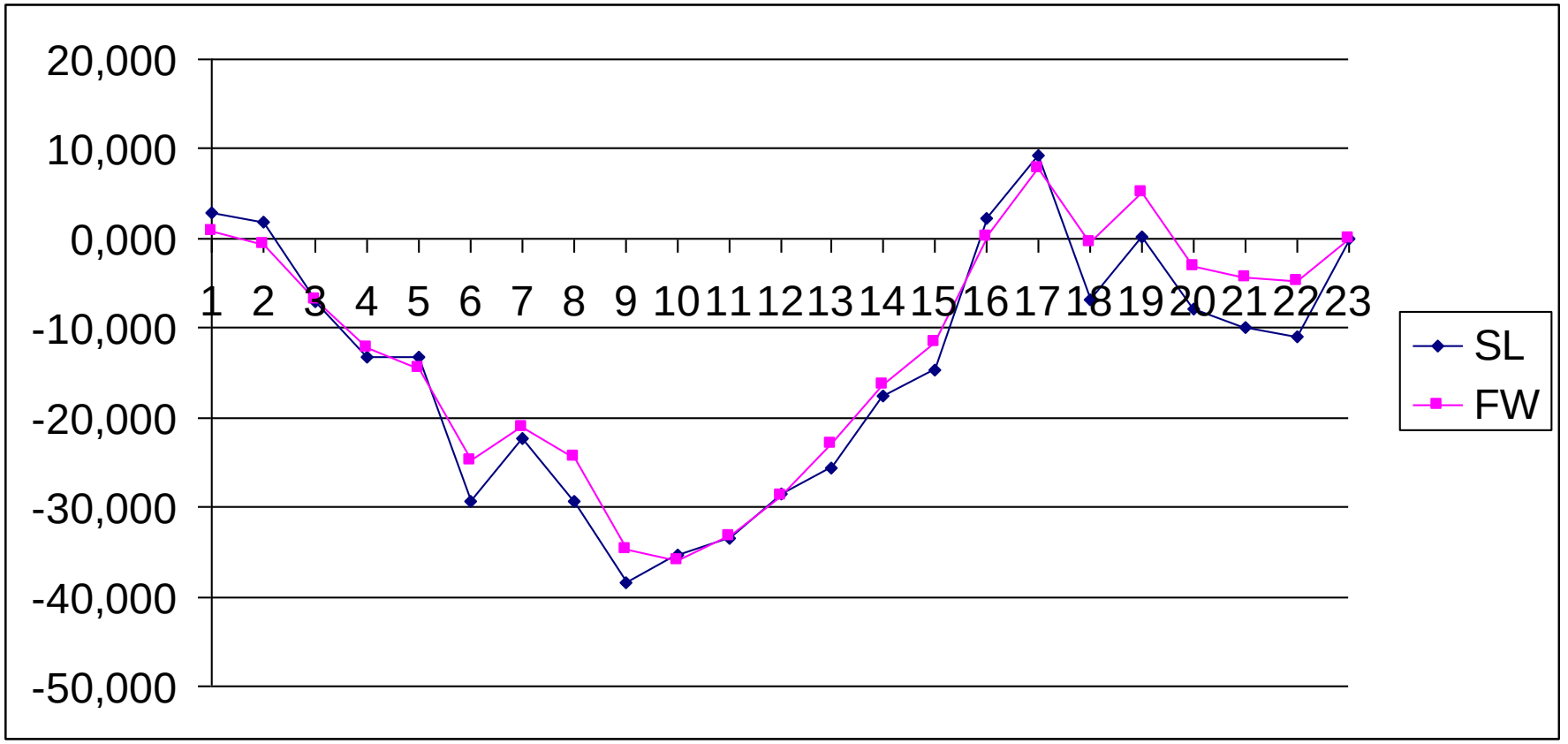
Αυτόματη Αναγνώριση Συγγραφέα

- **Λεξιλογικός πλούτος**
  - Yule's K
  - Λεξιλογική πυκνότητα
  - TTR
- **Μετρήσεις σε επίπεδο λέξης**
  - Μέσο μήκος λέξεων
  - Κατανομή του μήκους λέξεων
- **Μετρήσεις σε επίπεδο πρότασης**
  - Μέσο μήκος της πρότασης (σε λέξεις)
- **Σημεία στίξης**
- **Μεταβλητές «Διγλωσσίας»**
  - Ποικιλία των τριτόκλιτων καταλήξεων σε ης και –εως
  - Ποικιλία στη χρήση των αναφορικών αντωνυμιών «που» και «οποίος»
- **80 συχνότερες λειτουργικές λέξεις**

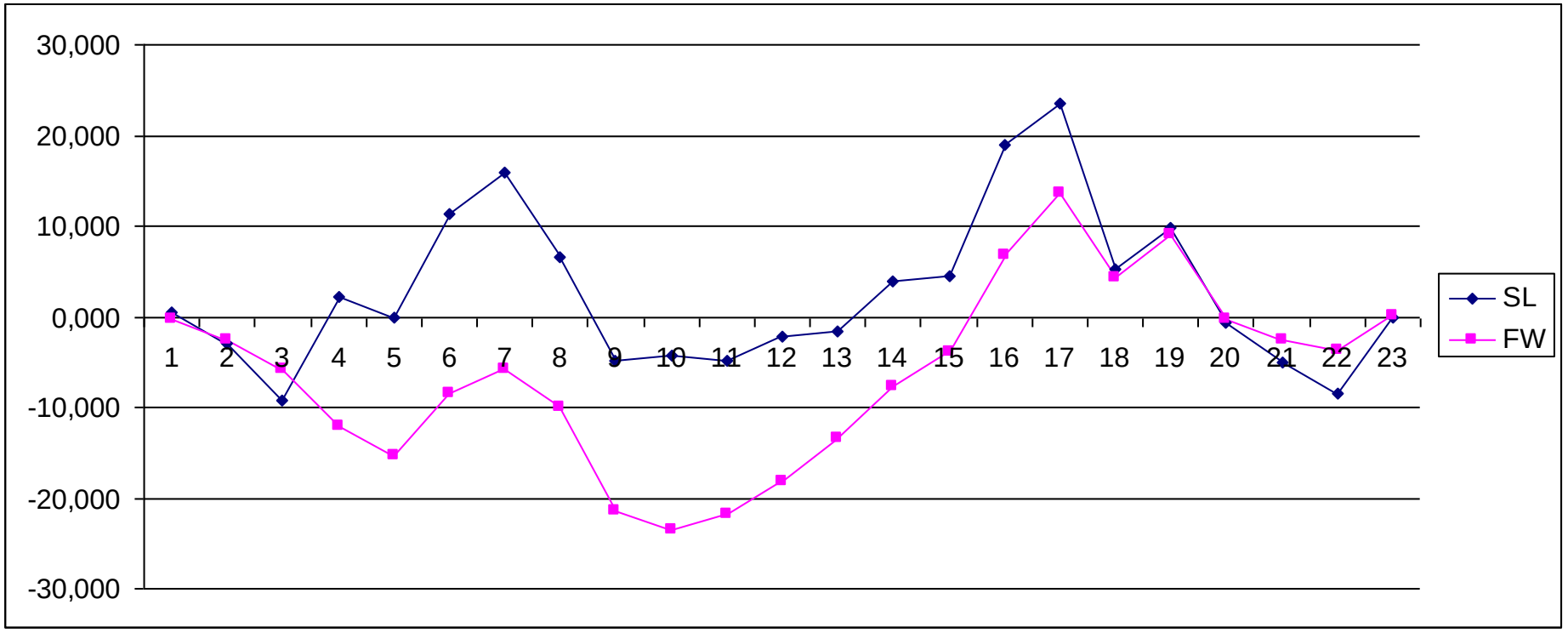
Αποτέλεσμα: Σε σύνολο 1200 κειμένων 93,6% ακρίβεια αναγνώρισης του συγγραφέα



Διάγραμμα CUSUM
1 συγγραφέας



Διάγραμμα CUSUM
2 συγγραφείς



Η σχέση υφομετρικών χαρακτηριστικών ενός κειμένου και της δυσκολίας κατανόησής του

Συμπεράσματα

- Η ποσοτικές μέθοδοι επιτρέπουν την θέαση της γλωσσικής χρήσης τόσο σε μικροσκοπικό όσο και σε μακροσκοπικό επίπεδο.
- Προσφέρουν ακριβείς ποσοτικές πληροφορίες που επιτρέπουν την μαθηματική περιγραφή των γλωσσικών φαινομένων.
- Επιτρέπουν τον εμπειρικό έλεγχο των γλωσσικών θεωριών.
- Λειτουργούν υποστηρικτικά σε ένα ευρύ φάσμα εφαρμογών γλωσσικής τεχνολογίας.

Αλλά...

- Τα αποτελέσματα εξαρτώνται άμεσα από την ποσοτική και ποιοτική σύσταση των ΗΣΚ που χρησιμοποιούμε.
- Η φύση των πληροφοριών που αντλούνται είναι πιθανοτική και όχι κατηγορική.
- Δεν καλύπτεται το σύνολο των γλωσσικών φαινομένων.

Βιβλιογραφικές αναφορές

- Μικρός, Γ. & Καραγιάννης, Γ. (υπό δημοσίευση). "Ποσοτική ανάλυση της χρήσης του κανόνα του τελικού -ν σε κείμενα της Νέας Ελληνικής". Στο *Γλωσσολογία.*
- Μικρός, Γ., Χατζηγεωργίου, Ν., Καραγιάννης, Γ. 2003. «Βασικά ποσοτικά μεγέθη στην γραπτή Νέα Ελληνική γλώσσα: η αξιοποίηση του ΕΘΕΓ στην ελληνική ποσοτική γλωσσολογία». *Proceedings of the Workshop "Text Processing for Modern Greek: From Symbolic to Statistical Aproaches"*, 20 Σεπτεμβρίου 2003, Ρέθυμνο, σσ. 23-37, ηλ. διαθέσιμο: http://www.philology.uoc.gr/conferences/6thICGL/ebook/ws/workshop@mikros.pdf
- Μικρός Γ. 2003. «Στατιστικές προσεγγίσεις στην αυτόματη κατηγοριοποίηση κειμένων της Νέας Ελληνικής: Μια πιλοτική αξιολόγηση υφομετρικών δεικτών και στατιστικών μεθόδων». Πρακτικά του *6ου Διεθνούς Συνεδρίου Ελληνικής Γλωσσολογίας,* 18-21 Σεπτεμβρίου 2003, Ρέθυμνο, ηλ. διαθέσιμο: Http://www.philology.uoc.gr/conferences/6thICGL/ebook/a/mikros.pdf
- Hatzigeorgiu, N., Mikros, G. & Carayannis, G. (2001). "Word length, word frequencies and Zipf's law in the Greek language". *Journal of Quantitative Linguistics*, Vol. 8, σσ. 175-185.
- Mikros, G. & Carayannis, G. 2000. "Modern Greek Corpus Taxonomy". *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC),* Vol.1, σσ. 129-134.
- Mikros, G., Hatzigeorgiu, N. & Carayannis, G. (2005). "Basic quantitative characteristics of the Modern Greek Language using the Hellenic National Corpus". *Journal of Quantitative Linguistics,* Vol. 12, σσ. 167-184.
- Mikros, G. (2005). "Quantitative linguistics in Greece: an overview". Στο Altmann, G., Kohler, R.& Piotrowski, R. (eds), *Quantitative Linguistics. An international handbook.* Berlin: Walter De Gruyter, σσ. 136-142.
- Mikros, G. (2006). "Authorship attribution in Modern Greek newswire corpora". Στο Uzuner, O., Argamon, S. & Karlgren, J. (eds), *Proceedings of the SIGIR 2006 Workshop on Directions in Computational Analysis of Stylistics in Text Retrieval*, Seattle, Washington, USA, August 10, 2006, σσ. 43-47, ηλ. διαθέσιμο: http://people.csail.mit.edu/ozlem/sigir_workshop_2006_proceedings.pdf

Τέλος Ενότητας

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αθηνών**» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σημειώματα

Σημείωμα Ιστορικού Εκδόσεων Έργου

Το παρόν έργο αποτελεί την έκδοση 1.0.

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.

Σημείωμα Χρήσης Έργων Τρίτων

"Η δομή και οργάνωση της παρουσίασης, καθώς και το υπόλοιπο περιεχόμενο, αποτελούν πνευματική ιδιοκτησία του συγγραφέα και του Πανεπιστημίου Αθηνών και διατίθενται με άδεια Creative Commons Αναφορά Μη Εμπορική Χρήση Παρόμοια Διανομή Έκδοση 4.0 ή μεταγενέστερη.

Οι εικόνες/σχήματα/διαγράμματα/φωτογραφίες που περιέχονται στην παρουσίαση αποτελούν πνευματική ιδιοκτησία τρίτων. Απαγορεύεται η αναπαραγωγή, αναδημοσίευση και διάθεσή τους στο κοινό με οποιονδήποτε τρόπο χωρίς τη λήψη άδειας από τους δικαιούχους. "