Η συμμετοχή του Πανεπιστημίου Αθηνών στην εθνική υποδομή CLARIN EL

Διονύσης Γούτσος

Κατηγορίες γλωσσικών πόρων

- Κατεξοχήν σώματα κειμένων
- Γλωσσάρια-λεξικά
- Ηλεκτρονικά περιοδικά & δημοσιεύσεις
- Αρχεία

Σώματα κειμένων

Τομέας Γλωσσολογίας

- ΣΕΚ www.sek.edu.gr
- ΣΚ αυθόρμητων προφορικών κειμένων Μπαμπινιώτη (χωρίς ηχητικά αρχεία)
- ΣΕΚ20: διαχρονικό ΣΚ 20ού αιώνα: υπό προετοιμασία
- ΣΚ ιδιωτικών επιστολών: υπό προετοιμασία
- ΣΚ αφασικού λόγου

Γλωσσάρια-Λεξικά

Τμήματα ξένων φιλολογιών

- Refranero Multilingue: πολύγλωσσο αρχείο παροιμιών με Ελληνικά

  http://cvc.cervantes.es/lengua/refranero/Default.aspx
- Γλωσσάρια γλωσσολογικών και άλλων όρων σε ηλεκτρονική μορφή
- Δίγλωσσο λεξικό ιδιωτισμικών εκφράσεων
- Δίγλωσσο γλωσσάρι λεξιλογίου ποδοσφαίρου
- Δίγλωσσο γλωσσάρι όρων για την κρίση

Ηλεκτρονικά περιοδικά και δημοσιεύσεις

Παραδείγματα:

- Γλωσσολογία/Glossologia
- Civitas Gentium
- Διαγλωσσικές Θεωρήσεις
- ΠΕΑΠ
- Αφορμές
- Παράβασις

Αρχεία Ι

Τμήμα ΦΠΨ

- Μουσείο και αρχείο παιδείας: ηχητικά αρχεία με συνεντεύξεις κ.λπ.

Τμήμα ΕΑΠΗ

- Δεδομένα από αφηγήσεις και συνομιλίες με παιδιά

Αρχεία ΙΙ

Τμήμα ΜΙΘΕ

- Εργαστήριο Ηλεκτρονικής Διαχείρισης Ιστορικών Αρχείων:

Ελληνομνήμων**:** Ψηφιακή βιβλιοθήκη (1600-1821)

Ανθέμιον: Αρχειακό Υλικό Ρωμαίικων Κοινοτήτων της Πόλης

Κάτοπτρο: Ιστορικό υλικό για το διάστημα 1453-1821

Αποδελτίωση περιοδικού 1863-1865

Ψηφιοποίηση εφημερίδας (1926-1952)

Σχολές του Γένους (βάση δεδομένων)

Λεξικό επιστημονικών όρων 17ου-19ου αιώνα

Αρχεία ΙΙΙ

Τμήμα Οδοντιατρικής

- Αρχειακό έντυπο υλικό σε μορφή flip book

Υπολογιστικό Κέντρο Βιβλιοθηκών

- Πέργαμος: αποθετήριο συλλογών
- Έφεσος: αποθετήριο διατριβών

Σώματα κειμένων της Ελληνικής Ι

	Ιστοσελίδα	Αριθμός λέξεων
ΕΘΕΓ	hnc.ilsp.gr	40.000.000
ΣΕΚ	sek.edu.gr	30.000.000
ΣΚ Πύλης	greeklanguage.gr	7.000.000
Προφορικό ΣΚ ΙΝΣ	corpus-ins.lit.auth.gr	1.700.000
Corpus of MG	web-corpora.net	35.700.000
+ GkWaC	sketchengine.co.uk	100.000.000

Σώματα κειμένων της Ελληνικής ΙΙ

* Πολύγλωσσα

	Ιστοσελίδα	Αριθμός λέξεων
JRC-Acquis	ec.europa.eu/jrc/en/language-technologies	
DGT-Acquis DGT-TM		(97.832.281)
Europarl	www.statmt.org/europarl	27.772.533
PANACEA	www.elra.info	600.000
MLCC	www.elra.info	
OPUS	opus.lingfil.uu.se/	
FREL	niobe.frl.auth.gr/	500.000

Σώματα κειμένων σε άλλες γλώσσες Ι

Αγγλικά:

- Oxford English Corpus: πάνω από 2,5 δισ.
- Cambridge English Corpus:
  - Written English πάνω από 1 δισ.
  - Spoken English 70 εκατ.
  - Business English 200 εκατ.
  - English for Specific Purposes
  - Academic English πάνω από 400 εκατ.
  - Learner English 200.000 γραπτά

Σώματα κειμένων σε άλλες γλώσσες II

corpus.byu.edu

corpora, size, queries = better resources, more insight

Overview
Corpora
Size, speed, queries
Insight into variation

History / updates
FAQ / questions
Researchers
Publications

Register
Modify profile

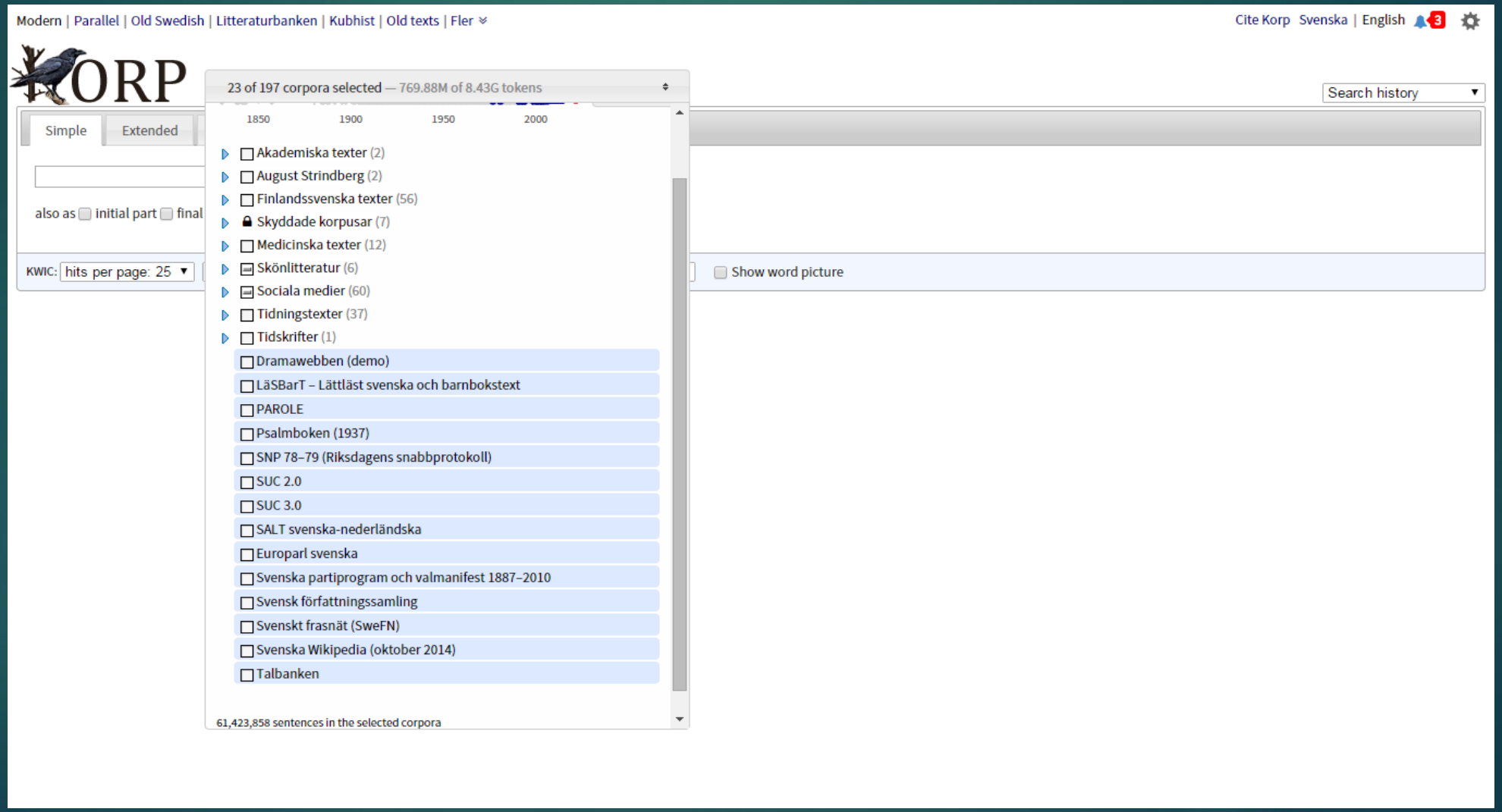
Related resources
Full-text data
Word frequency
Collocates
N-grams
WordAndPhrase
Academic vocabulary

Problems
Contact us

Created by Mark Davies, BYU. Overview, search types, researchers, publications, corpus-based resources.

English	# words	language/dialect	time period	compare
Global Web-Based English (GloWbE)	1.9 **billion**	20 countries	2012-13	
Corpus of Contemporary American English (COCA)	450 million	American	1990-2012	* * * * *
Corpus of Historical American English (COHA)	400 million	American	1810-2009	* *
TIME Magazine Corpus	100 million	American	1923-2006	
Corpus of American Soap Operas	100 million	American	2001-2012	*
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993	* *
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s	
Other languages				
Corpus del Español	100 million	Spanish	1200s-1900s	*
Corpus do Português	45 million	Portuguese	1300s-1900s	
N-grams				
Google Books: American English	155 **billion**	American	1500s-2000s	*
Google Books: British English	34 billion	British	1500s-2000s	
Google Books: One Million Books	89 billion	Am/Br	1500s-2000s	
Google Books: Spanish	45 billion	Spanish	1500s-2000s	

Σώματα κειμένων σε άλλες γλώσσες III



Σώματα κειμένων σε άλλες γλώσσες IV

- Εξειδικευμένα:
  - Διαχρονικά
  - Διαλεκτικά
  - Κοινωνιολέκτων
  - Πεδίων του λόγου
  - Μαθητικών κειμένων (learner corpora)
  - Επισημειωμένα για προσωδιακά φαινόμενα
  - Πολυμεσικά

Γλωσσικοί πόροι: ανάγκες

- Περισσότεροι, μεγαλύτεροι και εξειδικευμένοι πόροι

Κατευθύνσεις:

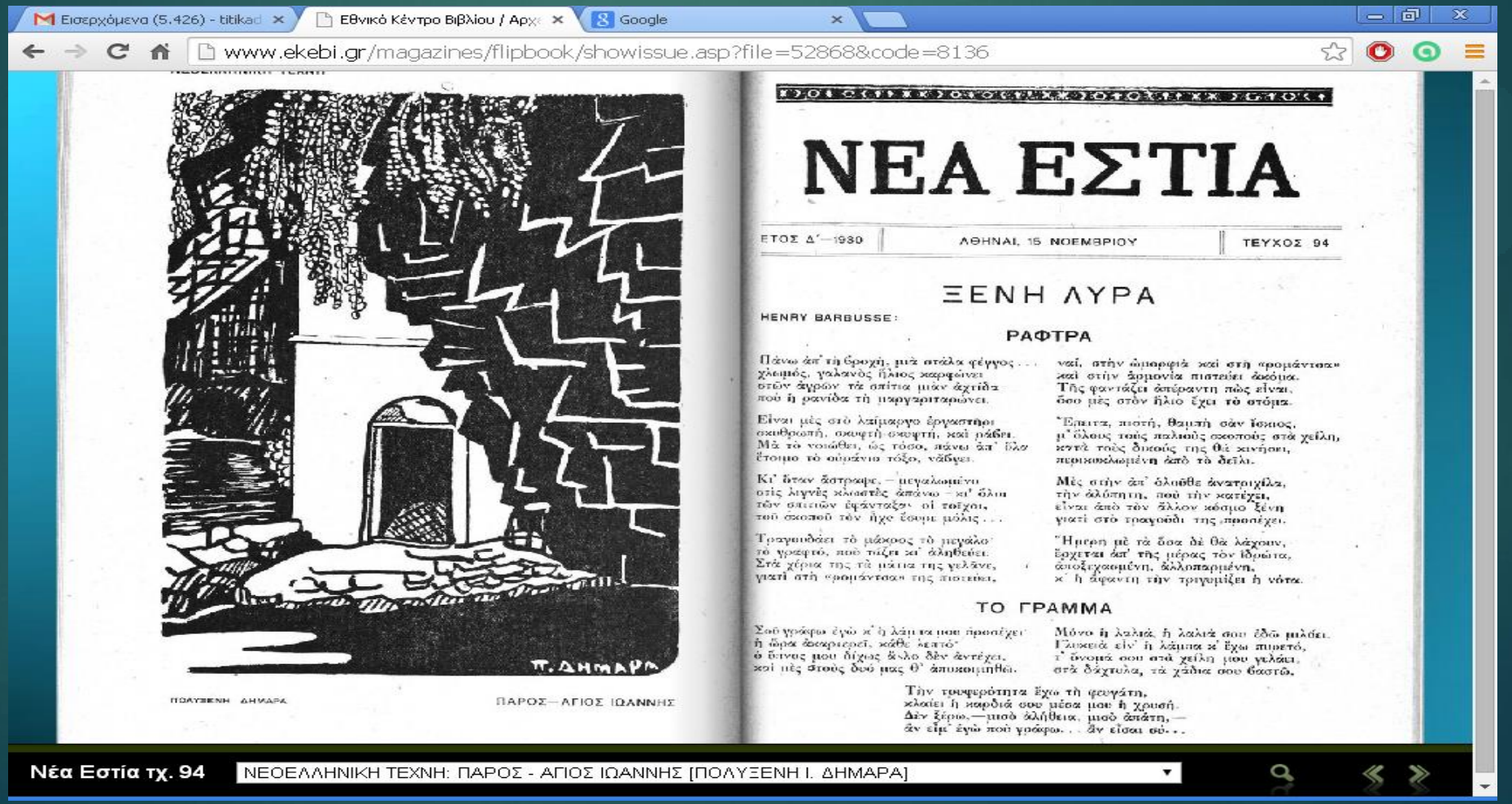
- Συλλογική προσπάθεια συγκέντρωσης και διαμοιρασμού δεδομένων
- Αλλαγή αντίληψης για τα ψηφιακά δεδομένα
- Ανάπτυξη εθνικής στρατηγικής

Γλωσσικοί πόροι: κατευθύνσεις Ι

- Συλλογική προσπάθεια συγκέντρωσης και διαμοιρασμού δεδομένων

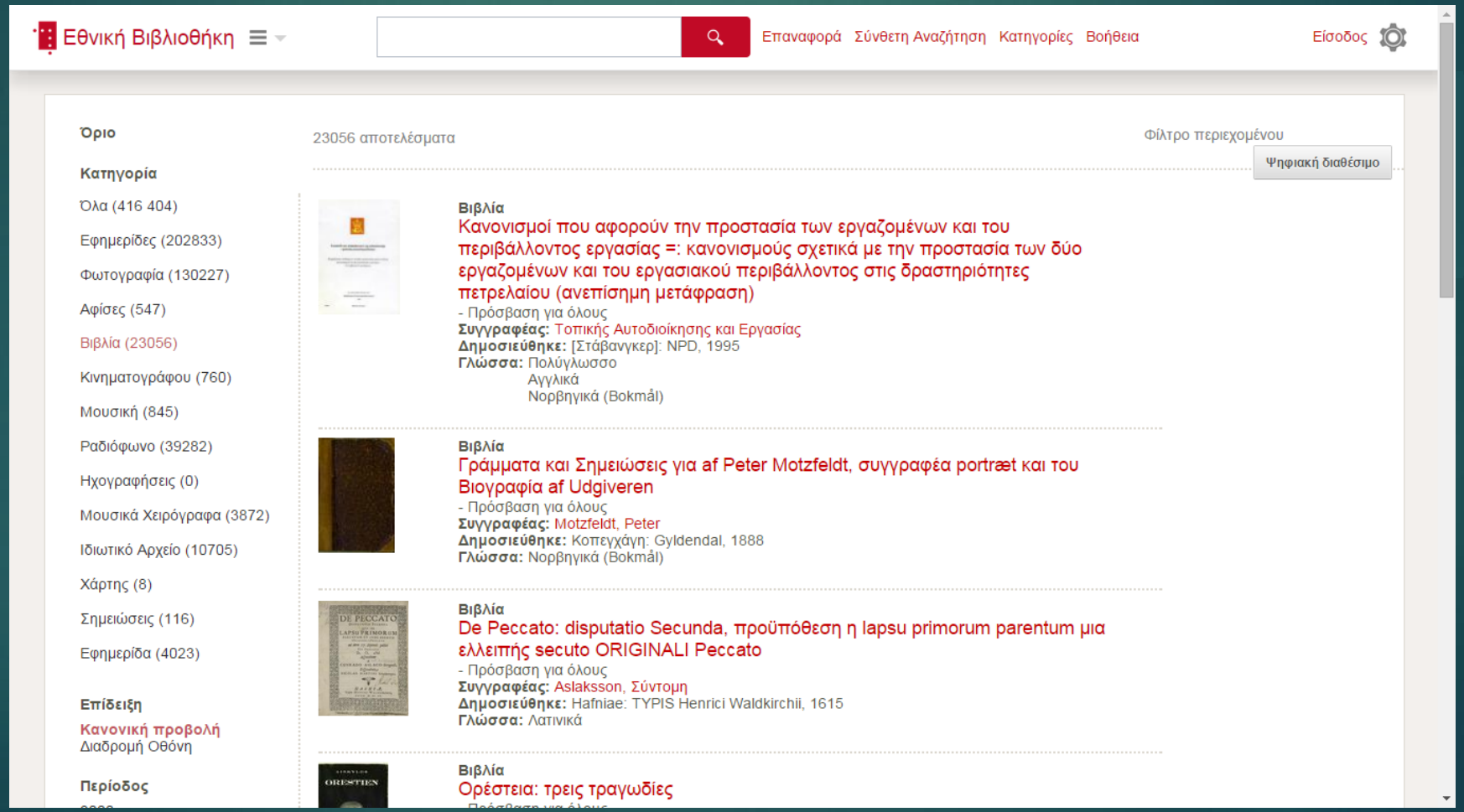
Γλωσσικοί πόροι: κατευθύνσεις ΙΙ

- Αλλαγή αντίληψης για τα ψηφιακά δεδομένα



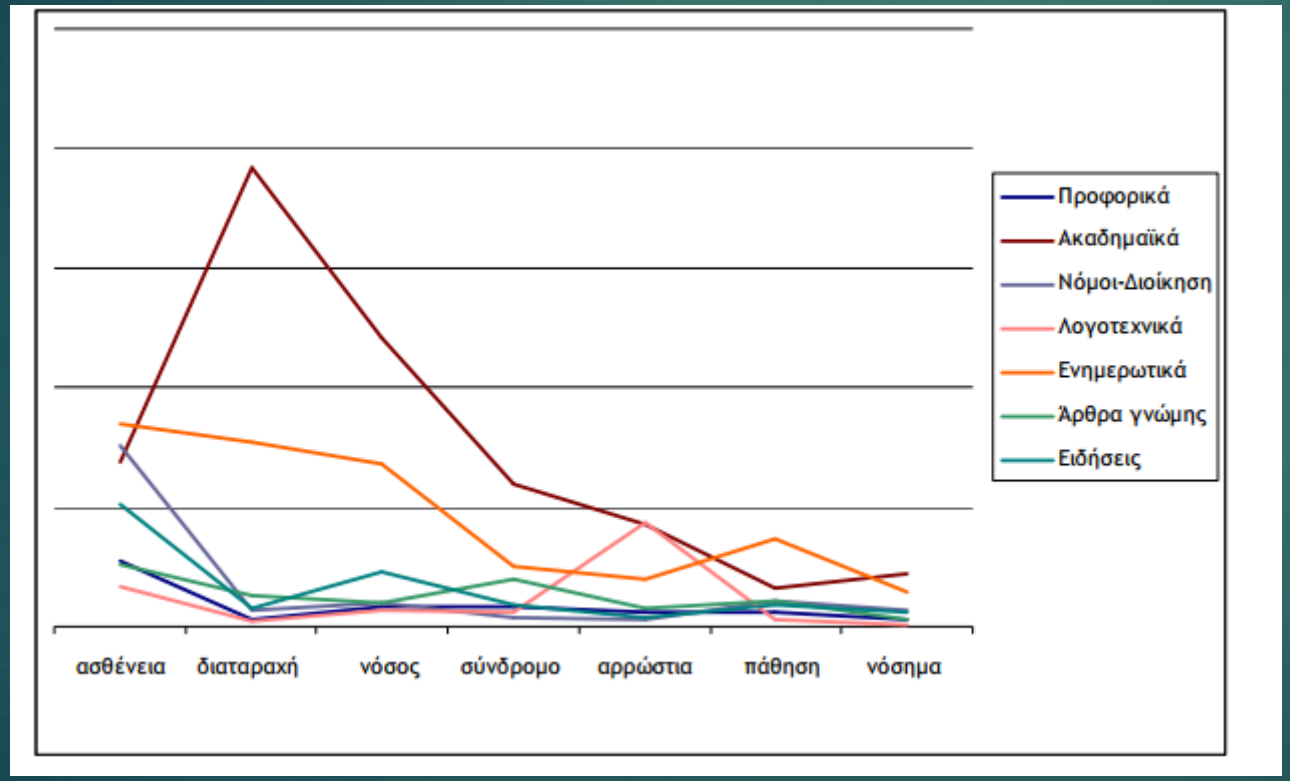
Γλωσσικοί πόροι: κατευθύνσεις ΙΙΙ

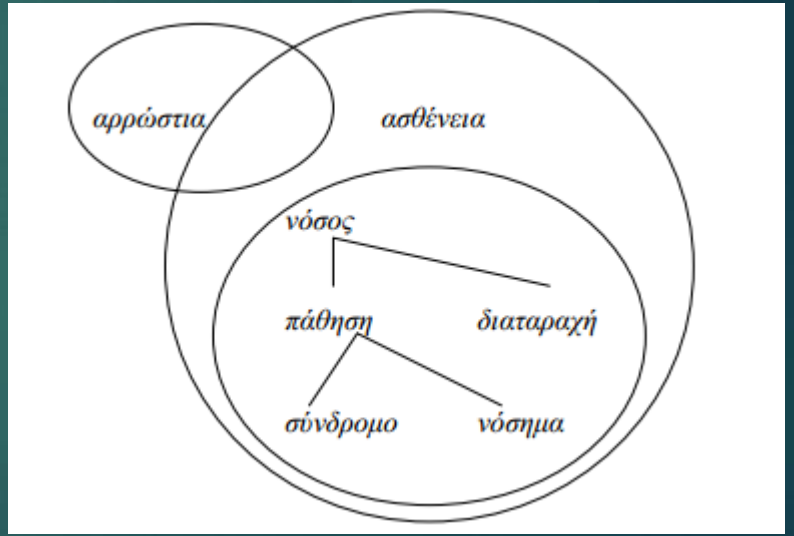
- Ανάπτυξη εθνικής στρατηγικής



Εφαρμογές Ι

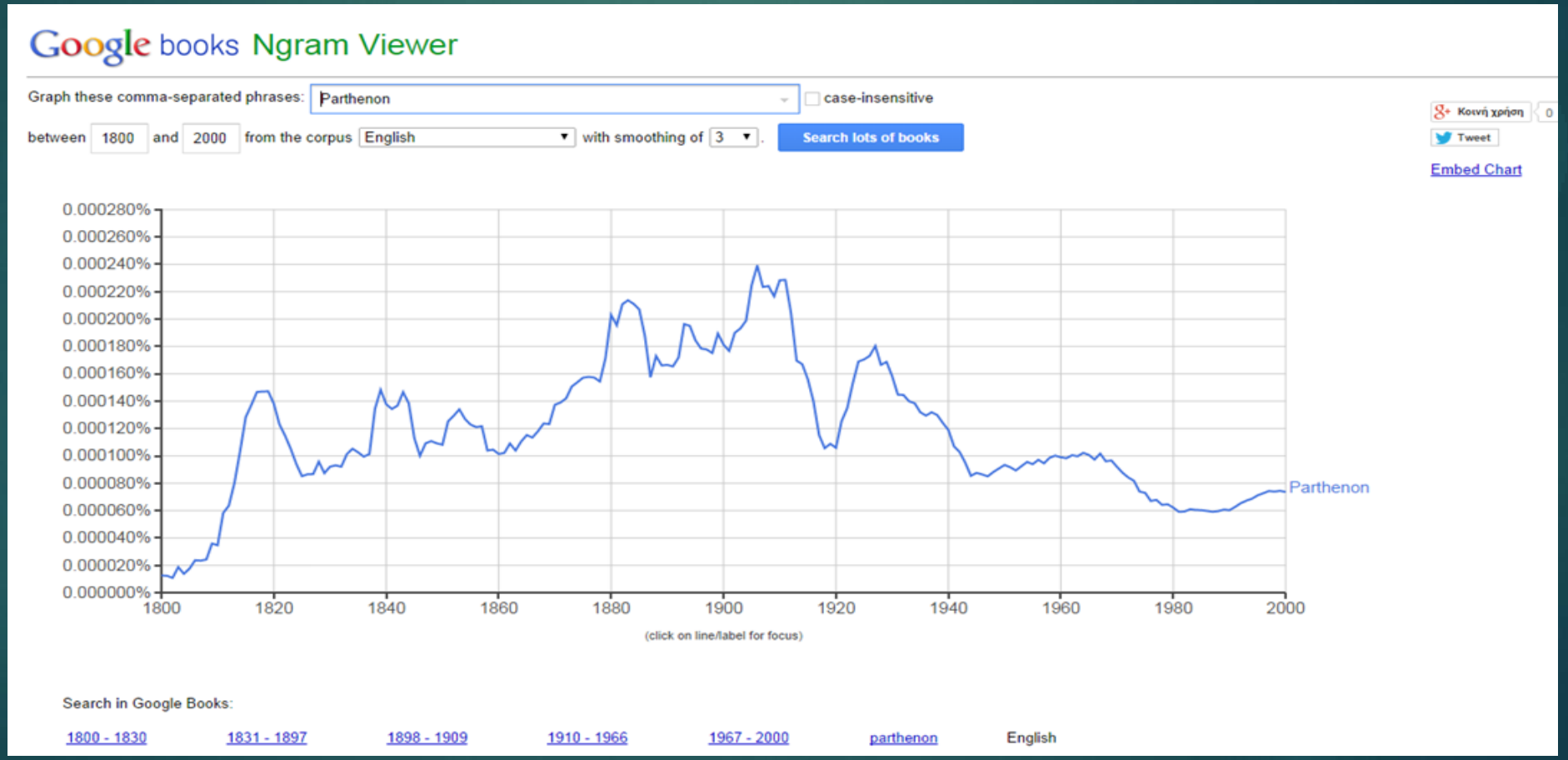
- σημασιολογική χαρτογράφηση πεδίου





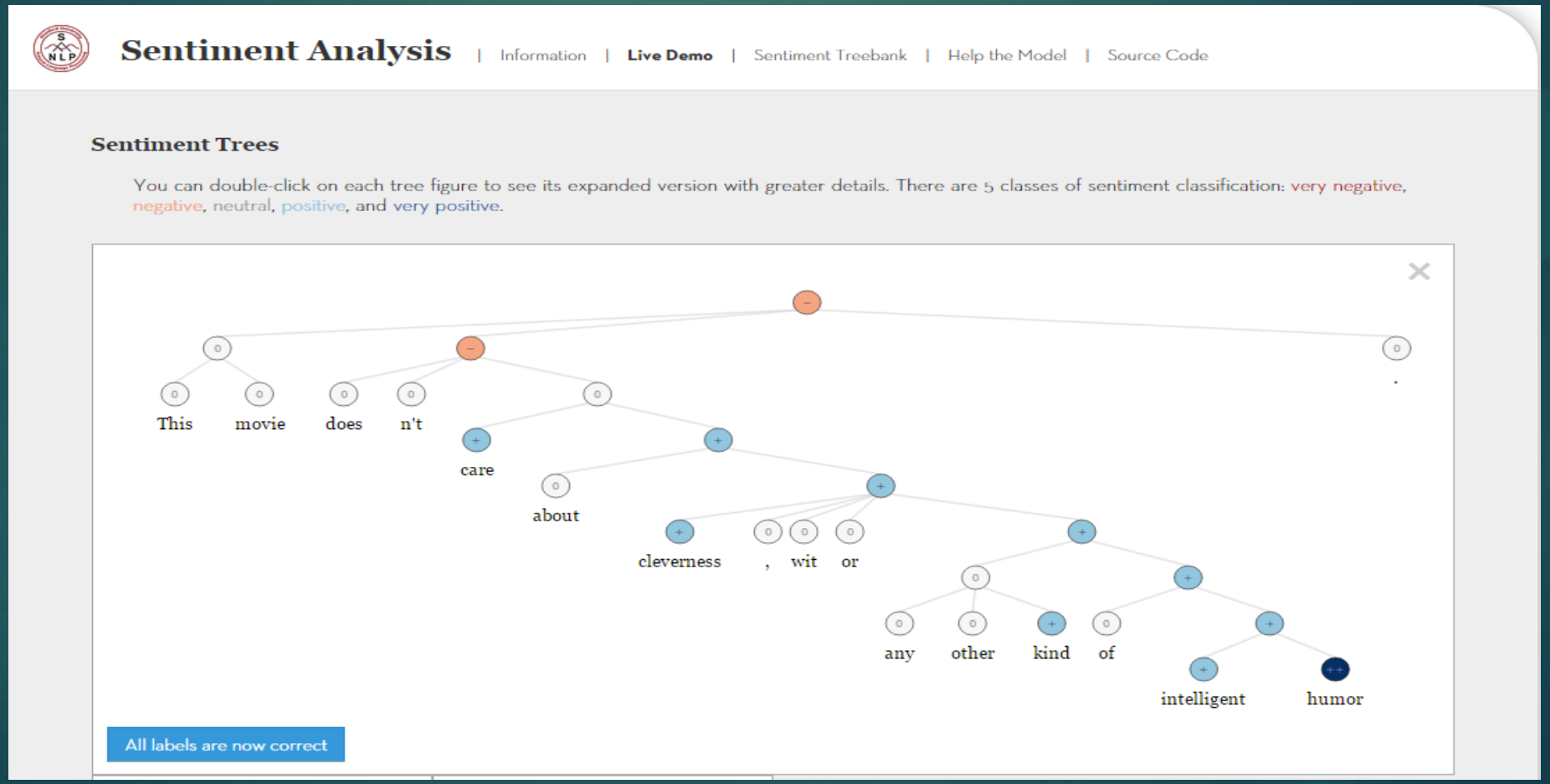
Εφαρμογές ΙΙ

- διαχρονική εξέλιξη θεμάτων/ιστορική πορεία



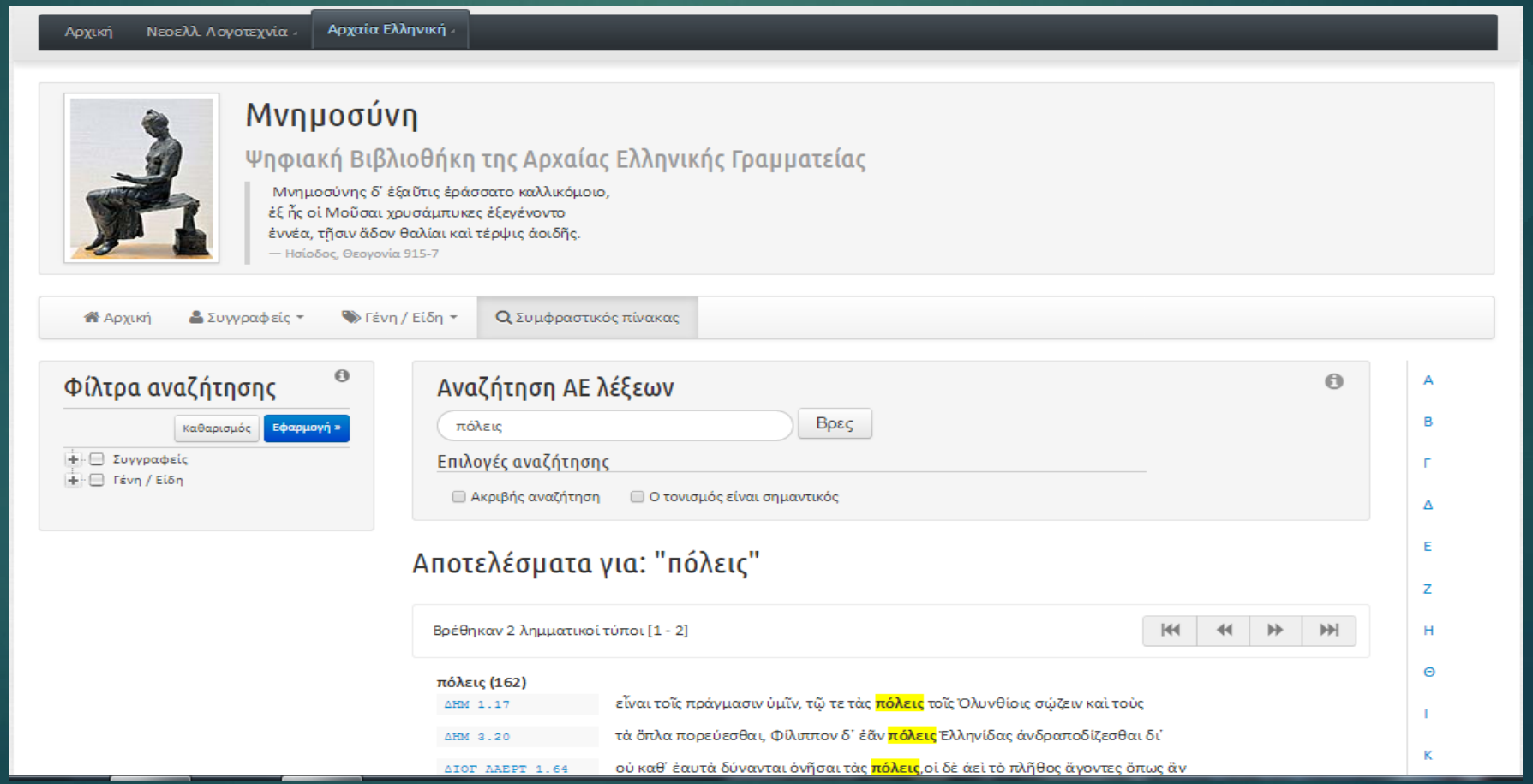
Εφαρμογές ΙΙΙ

- εξόρυξη γνωμών και ανάλυση τοποθέτησης/συναισθημάτων



Εφαρμογές IV

- εκπαιδευτικές εφαρμογές με στόχο τη διδασκαλία



Εφαρμογές V

- big data

Γλωσσικά εργαλεία: προοπτικές

- Εξειδικευμένα εργαλεία για παλαιότερες και διαλεκτικές ποικιλίες Ελληνικής π.χ. πολυτονικό OCR
- Ολοκληρωμένες πλατφόρμες ανάλυσης
- Διαθέσιμα εργαλεία σε φιλική μορφή για τους χρήστες