

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 II

Τμήμα ΜΙΘΕ

▶ Εργαστήριο Ηλεκτρονικής Διαχείρισης Ιστορικών Αρχείων:

Ελληνομνήμων: Ψηφιακή βιβλιοθήκη (1600-1821)

Ανθέμιον: Αρχειακό Υλικό Ρωμαϊκών 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 III

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 I

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

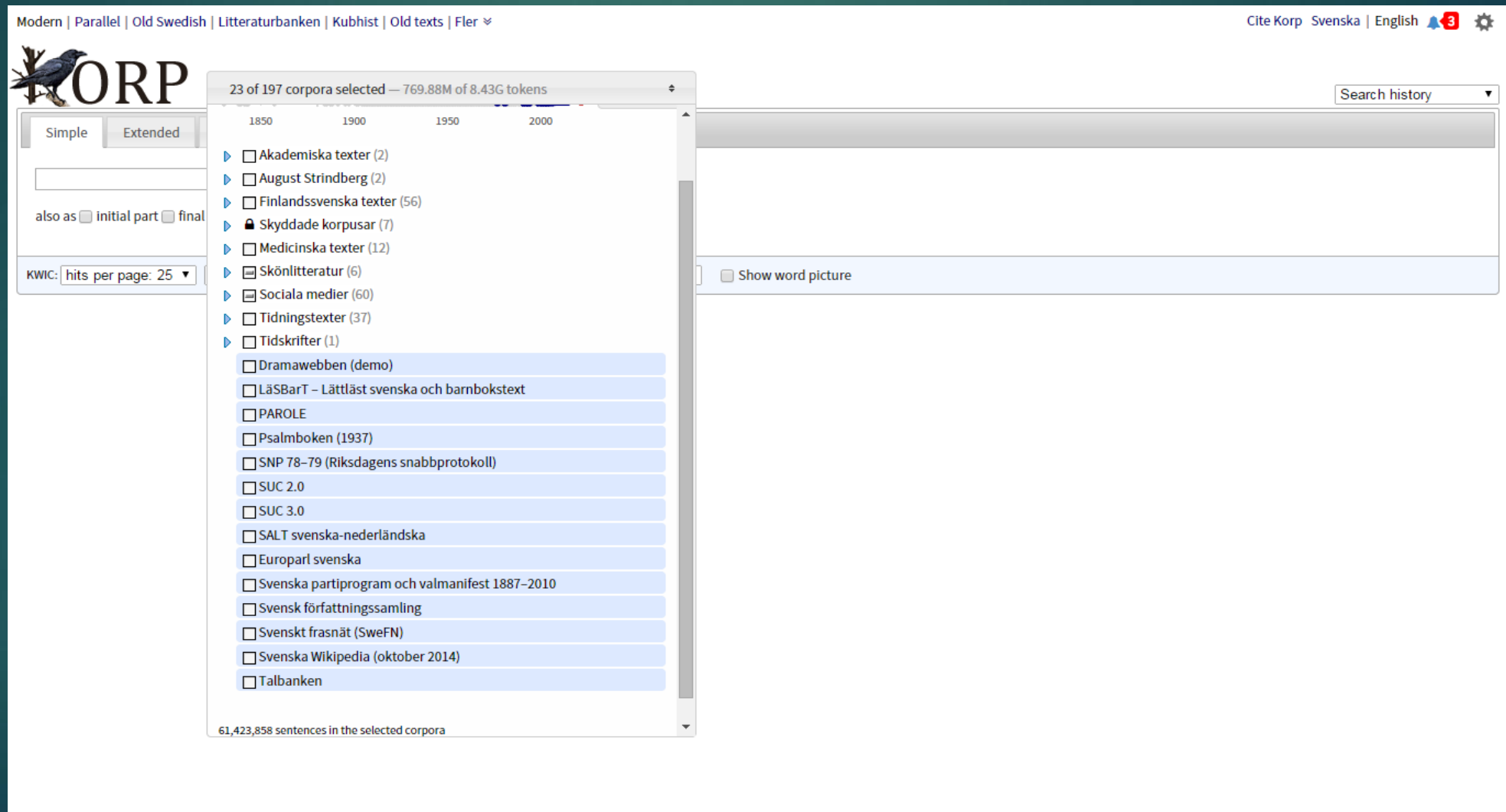
Created by Mark Davies, BYU. [Overview](#), [search types](#), [researchers](#), [publications](#), [corpus-based resources](#).

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

Modern | Parallel | Old Swedish | Litteraturbanken | Kubhist | Old texts | Fler

Cite Korp Svenska | English



KORP

23 of 197 corpora selected — 769.88M of 8.43G tokens

1850 1900 1950 2000

- Akademiska texter (2)
- August Strindberg (2)
- Finlandssvenska texter (56)
- Skyddade korpusar (7)
- Medicinska texter (12)
- Skönlitteratur (6)
- Sociala medier (60)
- Tidningstexter (37)
- Tidskrifter (1)
- Dramawebben (demo)
- LäSBarT – Lättläst svenska och barnbokstext
- PAROLE
- Psalmboken (1937)
- SNP 78–79 (Riksdagens snabbprotokoll)
- SUC 2.0
- SUC 3.0
- SALT svenska-nederländska
- Europarl svenska
- Svenska partiprogram och valmanifest 1887–2010
- Svensk författningssamling
- Svenskt frasnät (SweFN)
- Svenska Wikipedia (oktober 2014)
- Talbanken

61,423,858 sentences in the selected corpora

Simple Extended

also as initial part final

KWIC: hits per page: 25

Search history

Show word picture

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 I

- Συλλογική προσπάθεια συγκέντρωσης και διαμοιρασμού δεδομένων

Γλωσσικοί πόροι: κατευθύνσεις II

- ▶ Αλλαγή αντίληψης για τα ψηφιακά δεδομένα

www.ekebi.gr/magazines/flipbook/showissue.asp?file=52868&code=8136

ΝΕΑ ΕΣΤΙΑ

ΕΤΟΣ Δ'—1930 ΑΘΗΝΑΙ, 15 ΝΟΕΜΒΡΙΟΥ ΤΕΥΧΟΣ 94

ΞΕΝΗ ΛΥΡΑ

HENRY BARBUSSE:

ΡΑΦΤΡΑ

Πάνω απ' τη βροχή, μετ' στάλα φέγγος...
χλωμός, γαλανός ήλιος κερφαίνει
στον άγρον τό σπίτι μιν άχτίδα
πού ή ρανίδα τή μαργαριτωμένα.

Είναι μίς στο λαιμαργό έργαστήρι
σκυθρωπή, σκαμπή σκαμπή, καί ράβδι.
Μά τό νοιάζει, ως τόσο, πάλιν απ' ήλι
έτοιμο τό σούμνιο τόξο, νάβρει.

Κ' έπεν άστραψε, — μεγαλομύνη
στις λιγνές κλισιές άπάνω — κα' όλιν
τών σαισιών έφάνταξεν οι τοίχου,
τού σκοπού τόν ήχο έσφι μολίς...

Τραγουδάει τό μένος τό μεγάλο
τό φρεσίο, πού παίζει κα' άληθεία.
Στά χέρια της τό μάτι της γέλλει,
γιατί στή «ρομάντισα» της ποίαια,

ναί, στήν άμορφιά καί στή «ρομάντισα»
καί στήν άρμονία πισταίει άκόσμα.
Τής φαντάζει άπέραντη πώς είναι,
όσο μίς στον ήλιο έχει τό στόμα.

Έπειτα, πιστή, θαμνή σάν ίσκιος,
μ' άλλος τούς απλάς σκοπούς στή χείλη,
κατά τούς δικούς της θέ κινήσει,
πρικαικωμένη από τό δέλι.

Μίς στήν απ' άλλωθε άνατοιχίλα,
τήν άλόγητη, πού την κατέχει,
είναι από τόν άλλον κόσμο ξένη
γιατί στο τραγουδί της προσέχει.

Ήμερη με τό όσα δι θα λάχουν,
έρχεται απ' της μέρας τόν ιδρώτα,
αποξηλωμένη, άλλοπαρμένη,
κα' ή άφαντη την τραγουδεί ή νότα.

ΤΟ ΓΡΑΜΜΑ

Σού γράφο έγώ κα' ή λήν τα μου προσέχει
ή όρα άσκαρταί, καθε λεπτό,
ό τίνας μου ήγας άλλο δέν άντέχει,
καί μίς στους δύο μας θ' άποκοιμηθεί.

Μόνο ή λαλιά ή λαλιά σου έδω μάδαί.
Γλωσσά εν' ή λήνισα κα' έχω πιρετό,
τ' άνομα σου στή χείλη μου γλάσι,
σά δάχτυλα, τά χέρια σου θάστω.

Τήν τραφερότητα έχω τή φευγάτη,
κλειεί ή μορφή σου μέσα μου ή χρυσή.
Δέν ήρω, — μισό άλήθεια, μισό άπάτη, —
άν είμ' έγώ πού γράφο... άν είσαι εσύ...

ΠΟΛΥΞΕΝΗ ΔΗΜΑΡΑ ΠΑΡΟΣ—ΑΓΙΟΣ ΙΩΑΝΝΗΣ

Νέα Εστία τχ. 94 ΝΕΟΕΛΛΗΝΙΚΗ ΤΕΧΝΗ: ΠΑΡΟΣ - ΑΓΙΟΣ ΙΩΑΝΝΗΣ [ΠΟΛΥΞΕΝΗ Ι. ΔΗΜΑΡΑ]

Γλωσσικοί πόροι: κατευθύνσεις III

▶ Ανάπτυξη εθνικής στρατηγικής

The screenshot displays the National Library of Greece website interface. At the top, there is a navigation bar with the logo and name 'Εθνική Βιβλιοθήκη', a search bar, and links for 'Επανάφορα', 'Σύνθετη Αναζήτηση', 'Κατηγορίες', 'Βοήθεια', 'Είσοδος', and a settings icon. Below the navigation bar, the main content area shows search results for '23056 αποτελέσματα'. On the left, there is a sidebar with filters for 'Όριο', 'Κατηγορία', and 'Επίδειξη'. The 'Κατηγορία' filter is expanded, showing various categories with their respective counts. The main content area displays three book entries, each with a thumbnail image, the title, author, and publication details.

Εθνική Βιβλιοθήκη ☰ 🔍 [Επανάφορα](#) [Σύνθετη Αναζήτηση](#) [Κατηγορίες](#) [Βοήθεια](#) [Είσοδος](#) ⚙️

Όριο 23056 αποτελέσματα 🔍 Ψηφιακή διαθέσιμο

Κατηγορία

- Όλα (416 404)
- Εφημερίδες (202833)
- Φωτογραφία (130227)
- Αφίσες (547)
- Βιβλία (23056)**
- Κινηματογράφου (760)
- Μουσική (845)
- Ραδιόφωνο (39282)
- Ηχογραφήσεις (0)
- Μουσικά Χειρόγραφα (3872)
- Ιδιωτικό Αρχείο (10705)
- Χάρτης (8)
- Σημειώσεις (116)
- Εφημερίδα (4023)

Επίδειξη

- Κανονική προβολή**
- Διαδρομή Οθόνη

Περίοδος

Βιβλία
Κανονισμοί που αφορούν την προστασία των εργαζομένων και του περιβάλλοντος εργασίας =: κανονισμούς σχετικά με την προστασία των δύο εργαζομένων και του εργασιακού περιβάλλοντος στις δραστηριότητες πετρελαίου (ανεπίσημη μετάφραση)
- Πρόσβαση για όλους
Συγγραφέας: Τοπικής Αυτοδιοίκησης και Εργασίας
Δημοσιεύθηκε: [Στάβανγκερ]: NPD, 1995
Γλώσσα: Πολύγλωσσο
Αγγλικά
Νορβηγικά (Bokmål)

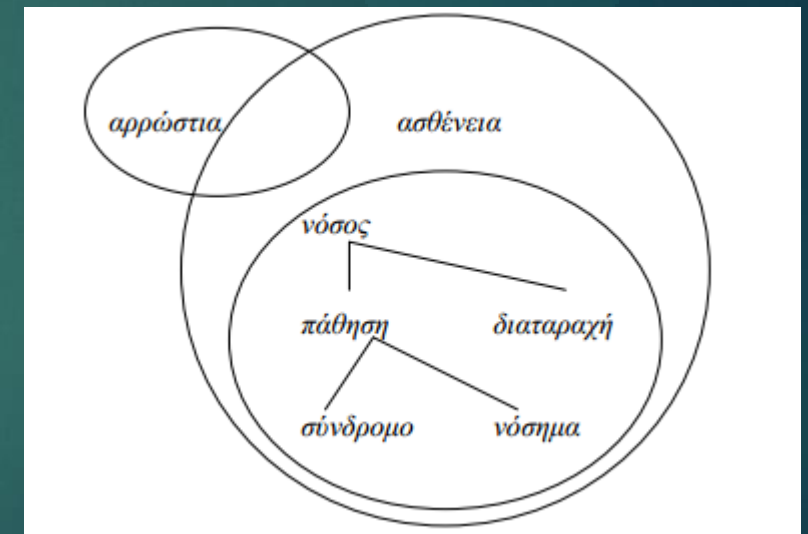
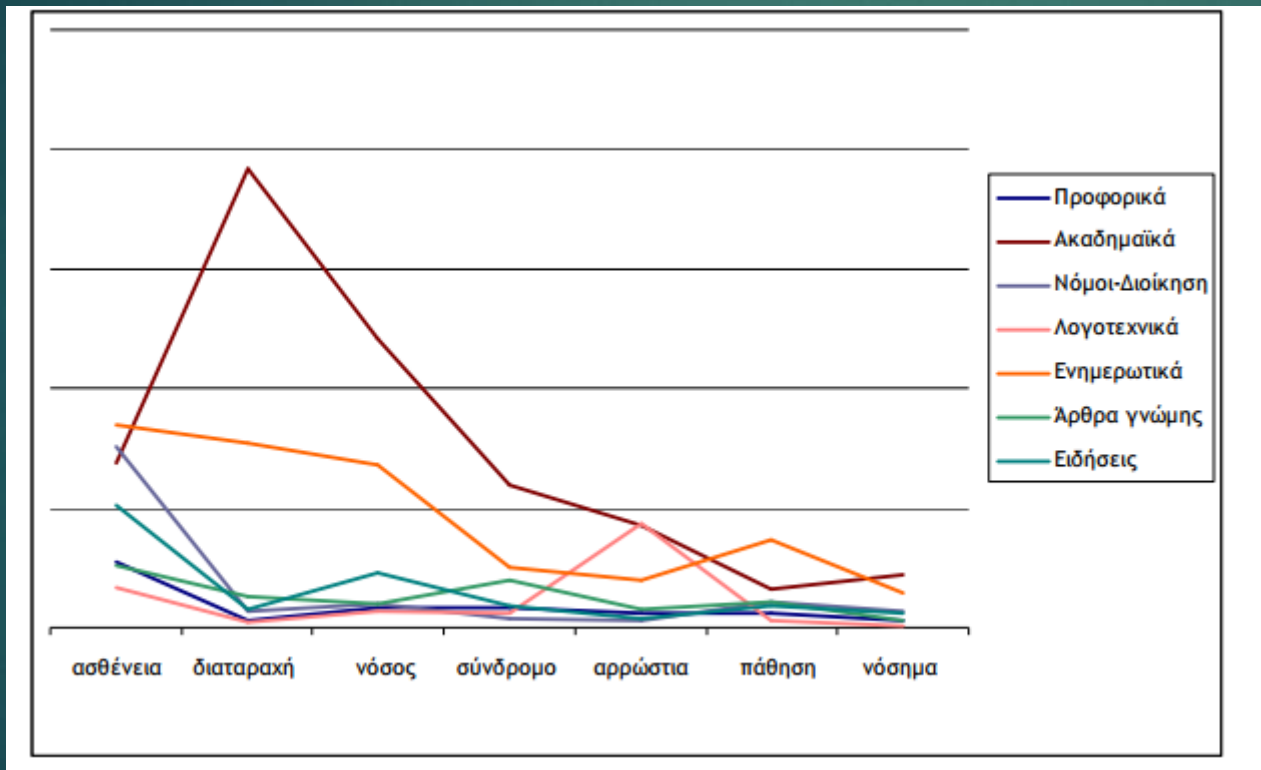
Βιβλία
Γράμματα και Σημειώσεις για af Peter Motzfeldt, συγγραφέα portræt και του Βιογραφία af Udgiveren
- Πρόσβαση για όλους
Συγγραφέας: Motzfeldt, Peter
Δημοσιεύθηκε: Κοπεγχάγη: Gyldendal, 1888
Γλώσσα: Νορβηγικά (Bokmål)

Βιβλία
De Peccato: disputatio Secunda, προϋπόθεση η lapsu primorum parentum μια ελλειπής secuto ORIGINALI Peccato
- Πρόσβαση για όλους
Συγγραφέας: Aslaksson, Σύντομη
Δημοσιεύθηκε: Hafniae: TYPIS Henrici Waldkirchii, 1615
Γλώσσα: Λατινικά

Βιβλία
Ορέστεια: τρεις τραγωδίες
- Πρόσβαση για όλους

Εφαρμογές I

- σημασιολογική χαρτογράφηση πεδίου




Εφαρμογές II

- διαχρονική εξέλιξη θεμάτων/ιστορική πορεία



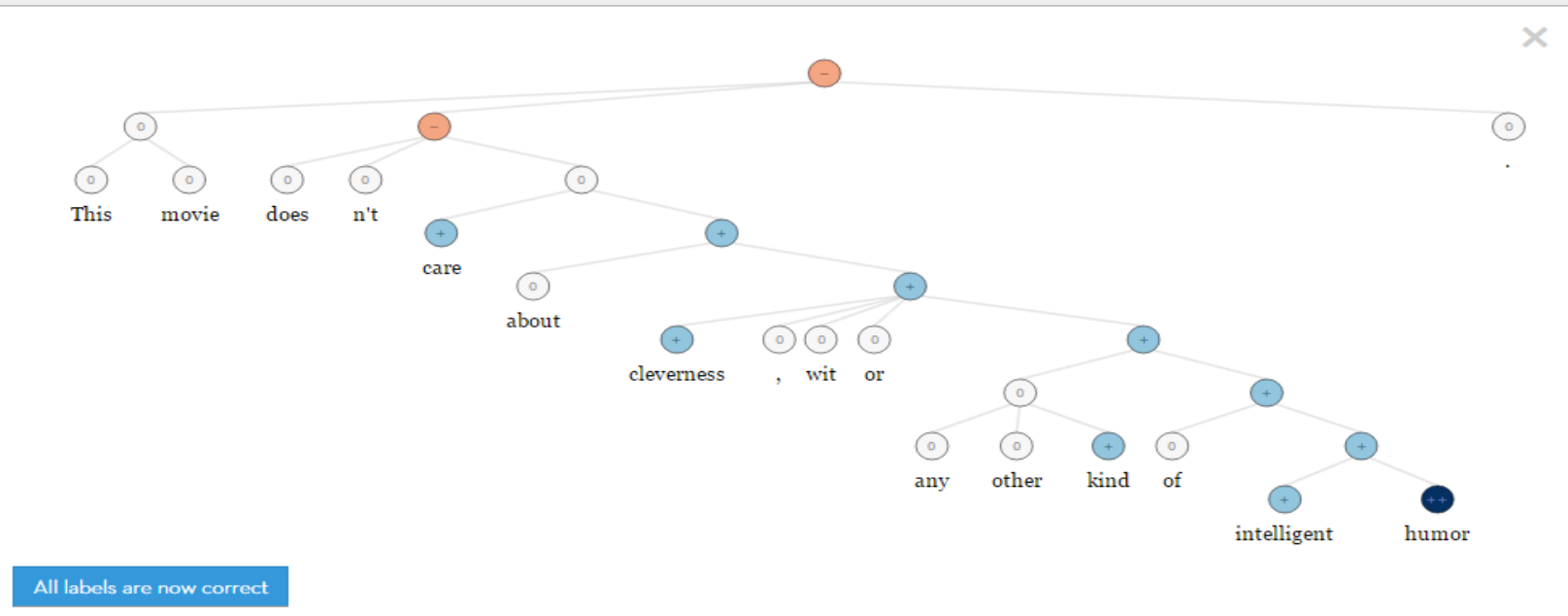
Εφαρμογές III

- εξόρυξη γνώμων και ανάλυση τοποθέτησης/συναισθημάτων

 **Sentiment Analysis** | [Information](#) | [Live Demo](#) | [Sentiment Treebank](#) | [Help the Model](#) | [Source Code](#)

Sentiment Trees

You can double-click on each tree figure to see its expanded version with greater details. There are 5 classes of sentiment classification: **very negative**, **negative**, **neutral**, **positive**, and **very positive**.



All labels are now correct

Εφαρμογές IV

- εκπαιδευτικές εφαρμογές με στόχο τη διδασκαλία

The screenshot displays the 'Μνημοσύνη' (Mnemosyne) digital library interface. At the top, there are navigation tabs for 'Αρχική', 'Νεοελλ. Λογοτεχνία', and 'Αρχαία Ελληνική'. The main header features a statue of Mnemosyne and the title 'Μνημοσύνη Ψηφιακή Βιβλιοθήκη της Αρχαίας Ελληνικής Γραμματείας'. Below this, a quote from Hesiod's 'Theogonia' (915-7) is shown: 'Μνημοσύνης δ' ἔξαυτις ἐράσσατο καλλικόμοιο, ἐξ ἧς οἱ Μοῦσαι χρυσάμπυκες ἐξεγένοντο ἑννέα, τῆσιν ἄδον θαλία καὶ τέρψις ἀοιδῆς. — Ησίοδος, Θεογονία 915-7'. A navigation bar includes 'Αρχική', 'Συγγραφείς', 'Γένη / Είδη', and a search bar for 'Συμφραστικός πίνακας'. On the left, a 'Φίλτρα αναζήτησης' (Search filters) panel shows 'Καθαρισμός' (Cleaning) and 'Εφαρμογή' (Application) buttons, along with expandable sections for 'Συγγραφείς' and 'Γένη / Είδη'. The main search area, 'Αναζήτηση ΑΕ λέξεων', has a search input with 'πόλεις' and a 'Βρες' (Find) button. Below the search bar, there are checkboxes for 'Ακριβής αναζήτηση' and 'Ο τονισμός είναι σημαντικός'. The results section, 'Αποτελέσματα για: "πόλεις"', shows 'Βρέθηκαν 2 λημματικοί τύποι [1 - 2]'. A vertical alphabetical index on the right lists letters A through K. The first result is for 'πόλεις (162)' with a reference to 'ΔΗΜ 1.17' and the text 'εἶναι τοῖς πράγμασιν ὑμῖν, τῷ τε τὰς πόλεις τοῖς Ὀλυνθίοις σῶζεν καὶ τοὺς τὰ ὄπλα πορεύεσθαι, Φίλιππον δ' ἔαν πόλεις Ἑλληνίδας ἀνδραποδίζεσθαι δι' οὐ καθ' ἑαυτὰ δύνανται ὀνήσῃαι τὰς πόλεις, οἱ δὲ αἰεὶ τὸ πλῆθος ἄγοντες ὅπως ἂν'.

Εφαρμογές V

- big data



The screenshot shows the Google Translate web interface. At the top left is the Google logo, and at the top right is the user's email address, dgoutsos@gmail.com. The page title is "Μετάφραση" (Translation). Below the title, there are language selection buttons: "Αγγλικά", "Ελληνικά", "Γερμανικά", and "Αναγνώριση γλώσσας". A double-headed arrow icon indicates the direction of translation. The source language is set to "Ελληνικά" and the target language is "Γαλλικά". A blue button labeled "Μετάφραση" is visible. The input text in the source box is "data sets that are too large and complex to manipulate or interrogate with standard methods or tools". The output text in the target box is "ensembles de données qui sont trop vaste et complexe pour manipuler ou interrogent des méthodes ou des outils standards". At the bottom, there is a footer with the text "Πληκτρολογήστε κείμενο ή διεύθυνση ιστότοπου ή μεταφράστε ένα έγγραφο."

Google

dgoutsos@gmail.com

Μετάφραση

Αγγλικά Ελληνικά Γερμανικά Αναγνώριση γλώσσας

Ελληνικά Γαλλικά Ιταλικά

Μετάφραση

data sets that are too large and complex to manipulate or interrogate with standard methods or tools

ensembles de données qui sont trop vaste et complexe pour manipuler ou interrogent des méthodes ou des outils standards

Πληκτρολογήστε κείμενο ή διεύθυνση ιστότοπου ή μεταφράστε ένα έγγραφο.

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