

Πρόγραμμα «Αριστεία»
«Διαχρονικό σώμα ελληνικών κειμένων του 20ού αιώνα/
Diachronic corpus of Greek of the 20th century»

Κ.Α. Έρευνας: 70/3/11920

Ακρωνύμιο: Greek Corpus 20

Κωδικός «ΑΡΙΣΤΕΙΑ Ι» 2396

Επιστημονικός υπεύθυνος: Διονύσης Γούτσος

ΤΕΧΝΙΚΗ ΑΝΑΦΟΡΑ ΓΙΑ ΤΟΥΣ ΓΛΩΣΣΙΚΟΥΣ ΠΟΡΟΥΣ
ΣΤΑ ΕΛΛΗΝΙΚΑ

Στη διαδικτυακή επικοινωνία για τους γλωσσικούς πόρους της Ελληνικής διαμορφώθηκε ο τελικός κατάλογος με τους γλωσσικούς πόρους της Ελληνικής, ο οποίος παρατίθεται πιο κάτω με τη μορφή πινάκων:

Βασικά μονόγλωσσα σώματα κειμένων Ελληνικής

<i>Ονομασία</i>	<i>Ιστοσελίδα</i>	<i>Αριθμός λέξεων</i>
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

Βασικά πολύγλωσσα σώματα κειμένων που περιλαμβάνουν Ελληνικά

<i>Ονομασία</i>	<i>Ιστοσελίδα</i>	<i>Αριθμός λέξεων</i>
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

Με βάση τα παραπάνω στοιχεία διαπιστώθηκε ότι οι πόροι της Ελληνικής είναι ιδιαίτερα περιορισμένοι, καθώς οι βασικές πηγές φθάνουν προσθετικά τα 100 εκατομμύρια λέξεις μόνο, ενώ τα πολύγλωσσα σώματα κειμένων που δημιουργήθηκαν ειδικά για τα Ελληνικά δεν ξεπερνούν τις 500.000 λέξεις.



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο



ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



ΕΣΠΑ
2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Επιπλέον, από ποιοτική άποψη τα βασικά μονόγλωσσα σώματα κειμένων της Ελληνικής είναι κυρίως:

- συγχρονικά 20ού αιώνα
- γραπτά
- πρότυπης γλώσσας.

Επομένως, προκύπτει η ανάγκη για περισσότερα, μεγαλύτερα και εξειδικευμένα Σώματα Κειμένων με έμφαση σε σώματα κειμένων:

- διαχρονικά
- διαλεκτικά
- κοινωνιολέκτων
- πεδίων του λόγου
- μαθητικών κειμένων (learner corpora)
- επισημειωμένα για προσωδιακά φαινόμενα
- πολυμεσικά.

Λεπτομερέστερη παρουσίαση των πορισμάτων για τους γλωσσικούς πόρους της Ελληνικής γίνεται στο παραδοτέο 7.2. του έργου.



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο



ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ