

17/1/2014



National and Kapodistrian University of Athens
Department of Linguistics

Greek corpus building and analysis:
The story so far and what is to follow

Dionysis Goutsos

Greek corpus building and analysis: The story so far and what is to follow

- *Greek corpus building*

1. Peculiarities of Greek

2. Phases of development

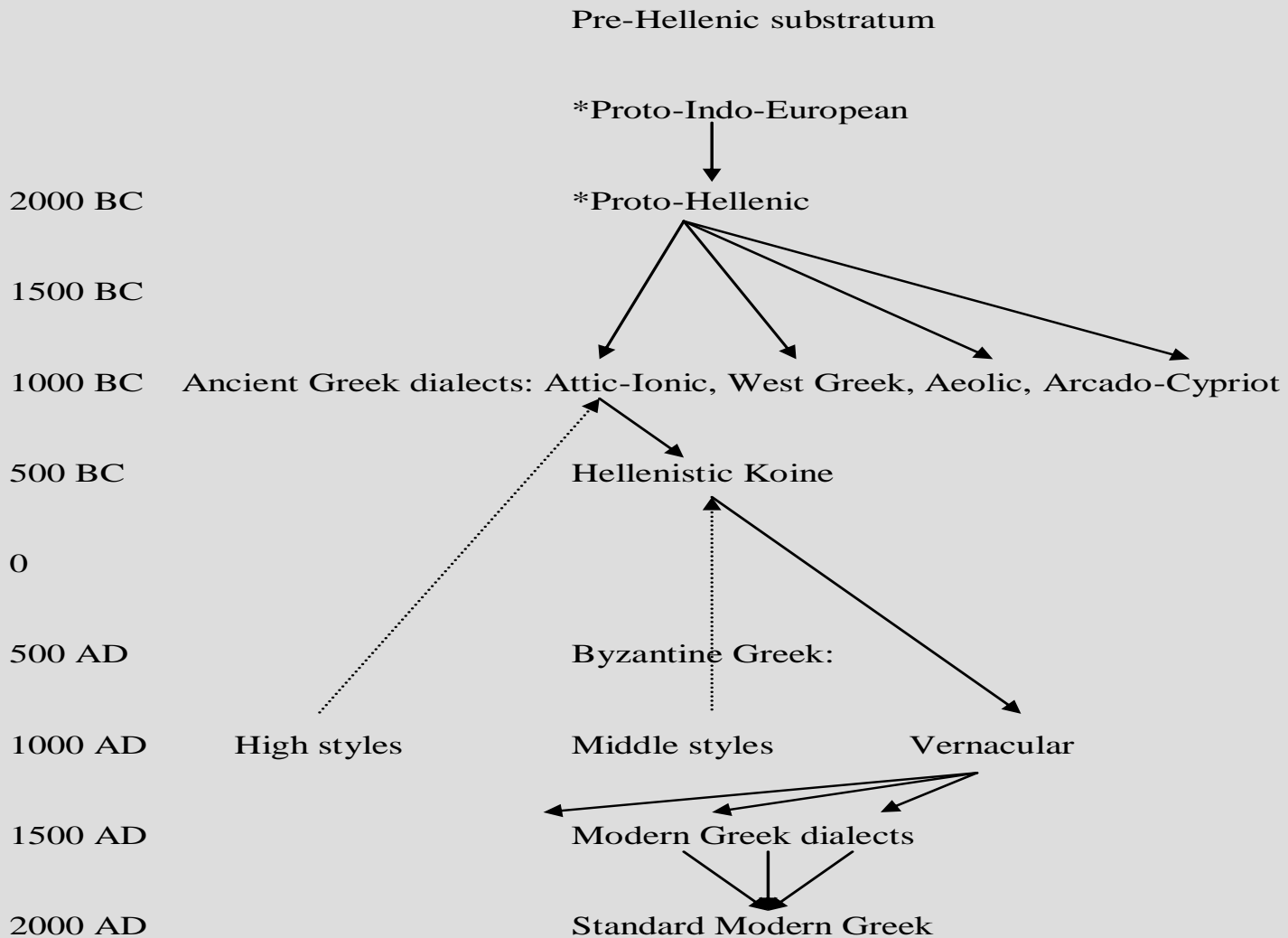
- *Corpus-based analysis of Greek*

3. Major findings on Greek

4. Major needs and prospects

5. Greek and corpus-based translation studies

1. Peculiarities of Greek



1. Peculiarities of Greek

- long historical tradition:
 - ▶ multiple continuities
 - ▶ genre and language varieties
- non-Western alphabet:
 - ▶ from ASCII to Unicode
 - ▶ 'monotonic' and 'polytonic' varieties
- aversion to empiricism
 - ▶ modern reference works (dictionaries and grammars: post-1998)
 - ▶ language resource development (lack of coordination and open-source tools)

2. Phases of development

Renouf (2007): English language corpora

- 1960s-1970s: the one-million word Small Corpus: standard, sampled: *LOB, Brown corpus*
- 1980s: the multi-million word Large Corpus: super-corpus: *Bank of English, BNC*
- 1990s: the 'Modern Diachronic' Corpus: dynamic, open-ended, chronological data flow: *FLOB, Frown*
- 1998-: the Web as corpus: cyber-corpus

2. Phases of development

Greek corpora

- 1980s: literary corpora: *Erotokritos* (1986), Makriyannis (1992)

Γ159 “Νένα μου, λέγει ἡ Ἀρετή, φρόνιμα δασκαλεύεις,
Γ171 “Παιδί μου, λέγει ἡ νένα της, σφάνουσι τὰ λογιάζεις,
Γ299 “Νένα μου, λέγει ἡ Ἀρετή, ἴντά ’ν’ τὰ δασκαλεύεις;
E613 “Νένα, τῇ λέγει ἡ Ἀρετή, τὸ γίνηκεν ἐγίνη
E1133 “Ἄμε πέ, λέγει ἡ Ἀρετή, γλήγορα τοῦ κυροῦ μου

0908	λύσει/ καὶ σὰν ὄντας κρατεῖς	κερὶ κι ἄνεμος σοῦ τὸ σβήσει// ἐκεῖνος ὁποῦ
1145	[οὐδὲ καπνίσματα μποροῦν οὐδὲ	κερὶα οὐδὲ θρόνοι// τὸν κόσμον ὅλο
0415	νὰ μ’ ἀφήσει/ ἐπά ’ν τ’ ἀφτούμενο	κερὶ ποῦ μελετᾶς νὰ σβήσεις// καὶ τὸ κορμί ὁποῦ
0408	μηδ’ εἶδα το ποτέ μου,/ μὰ ’ναν	κερὶν ἀφτούμενο ἐκράτουν κι ἤσβησέ μου//

(a) *Sacrifice of Abraham*

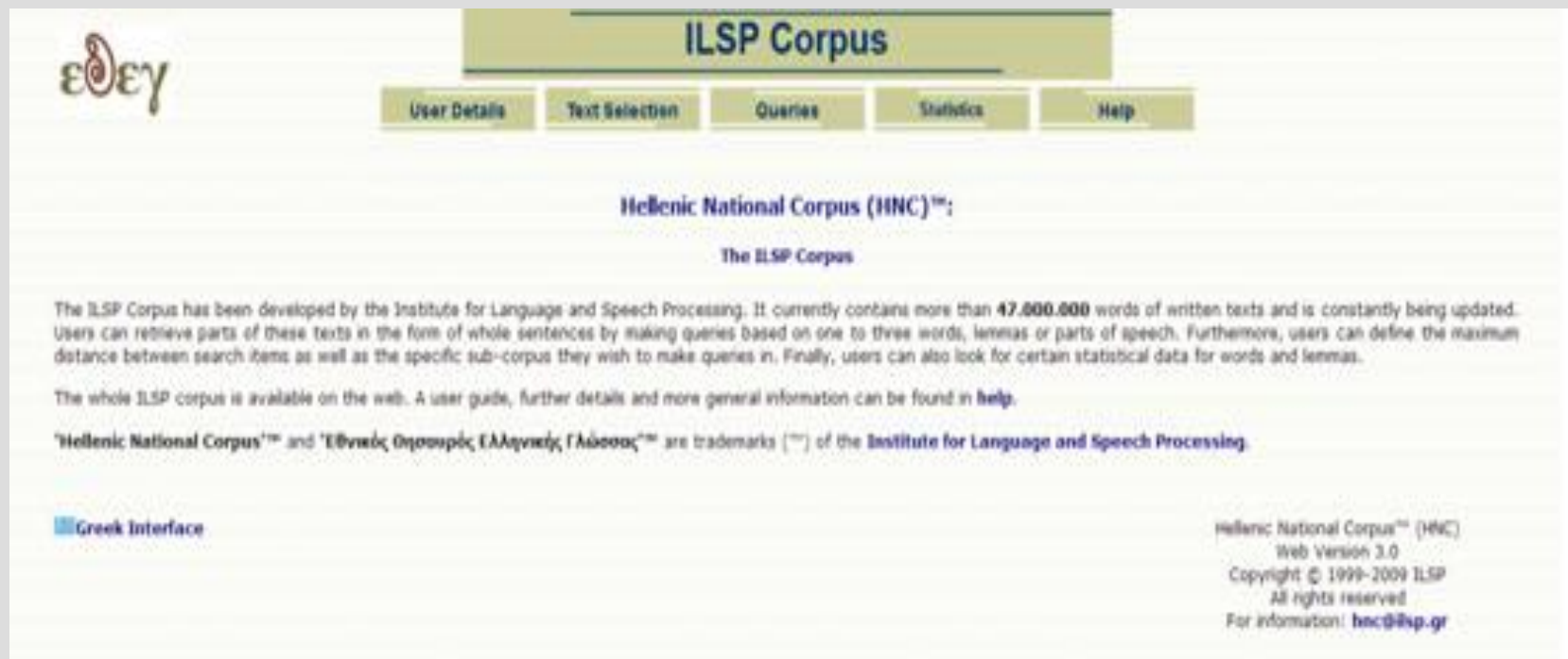
A0756 ΠΟΙ	κ’ ἐπόνει/ σὰν τὸ	κερὶ ἀνελίγωνε κ’ ἐφύρα σὰν τὸ χιόνι//
E0923 ΡΩΤ	ἀγάλια ἐχάνετο σὰν τὸ	κερὶ ὄντε σβήνη,/ ἤκλαψα κ’ ἐλυπήθηκα πολλὰ
Γ1102 ΦΡΟ	κι ὡσὰ φυσήξης τὸ	κερὶ ὁποῦ ’ψες, νὰ τὸ σβήσης// καὶ δὴ το καὶ
B2391 ΠΟΙ	κι οὐδὲ ψυχάρι οὐδὲ	κερὶ οὐδὲ φωτιά οὐδὲ γράμμα/ τοῦ ἐπόμεινε
B0757 ΠΟΙ	κι αὐτὸς στὴν κεφαλὴν ἕνα	κερὶ σβημένο/ τὸν Ἄνεμον ἀνάδια του ἤδειχνε
E0414 ΠΟΙ	τῶν ἀμματιῶν του ἢ λάμψη/	κερὶ σβηστὸν ἐφύλαγε καὶ πάγει ἐδὰ νὰ τ’ ἄψη//
B2177 ΠΟΙ	οἱ τρεῖς τῶς σ’ μιὰ μερὰ,	Κερί, Ψυχάρι, Ἀμάξι/ μὲ πεθυμιὰ ἀνιμένασι
Γ1396 ΡΩΤ	δὲ σ’ εἶδα ποτέ μου,/ μὰ ἕνα	κερὶν ἀφτούμενον ἐκράτουν κ’ ἤσβησέ μου//
Δ1763 ΠΟΙ	λαβώση// σὰ νὰ ’χεν εἶσται	κέρτινο, τέτοιας λογιῆς διαβαίνει/ στὸν κάμπο

(b) *Erotokritos*

2. Phases of development

Greek corpora

- 1990s-2000s: two 'super-corpora'
Hellenic National Corpus: <http://hnc.ilsp.gr>



The screenshot shows the ILSP Corpus website interface. At the top left is the logo of the Institute for Language and Speech Processing (ILSP), which consists of the Greek letters 'ΕΔΕΥ' in a stylized font. To the right of the logo is a green header bar with the text 'ILSP Corpus'. Below the header bar is a navigation menu with five buttons: 'User Details', 'Text Selection', 'Queries', 'Statistics', and 'Help'. The main content area features the title 'Hellenic National Corpus (HNC)™:' followed by the subtitle 'The ILSP Corpus'. Below this, there is a paragraph of text describing the corpus: 'The ILSP Corpus has been developed by the Institute for Language and Speech Processing. It currently contains more than 47,000,000 words of written texts and is constantly being updated. Users can retrieve parts of these texts in the form of whole sentences by making queries based on one to three words, lemmas or parts of speech. Furthermore, users can define the maximum distance between search items as well as the specific sub-corpus they wish to make queries in. Finally, users can also look for certain statistical data for words and lemmas.' Below this paragraph is another line of text: 'The whole ILSP corpus is available on the web. A user guide, further details and more general information can be found in [help](#).' At the bottom left, there is a link for 'Greek Interface'. At the bottom right, there is a footer with the following text: 'Hellenic National Corpus™ (HNC) Web Version 3.0 Copyright © 1999-2009 ILSP All rights reserved For information: hnc@ilsp.gr'.

2. Phases of development

- HNC: currently 47 million words
- lemmatized
- Problems with HNC:
 - restricted range of text types (newspaper articles)
 - no spoken data
 - restricted text classification (mostly 'other')
 - restricted availability
- > example of corpus building as separate from corpus analysis
- > need for a representative/authoritative corpus of Greek

2. Phases of development

Corpus of Greek Texts (CGT):
www.sek.edu.gr



πώς κι από ποῦ ἀκουμπάει τ'ὠμέγα στὸ ἄλφα

Σώμα Ἑλληνικῶν Κειμένων

Σύνδεση

Όνομα:

Συνθηματικό:

[Επιστροφή](#)
[Πατήστε ἕδωδε γιὰ βοήθησή μου...](#)

Σώμα Ἑλληνικῶν Κειμένων

Το Σώμα Ἑλληνικῶν Κειμένων (ΣΕΚ) δημιουργήθηκε με στόχο τὴ ψηφιοποίηση τῆς ἑλληνικῆς καὶ εἶναι τὸ πρῶτο ηλεκτρονικὸ σῶμα κειμένων που περιλαμβάνει ἑνα πλούσιό πηροφόρῳν καὶ κρατῶν κειμένων ὁδῶν τῆς σύγχρονης γλώσσας.

Το Σώμα Ἑλληνικῶν Κειμένων αποτελεί προϊόν τῆς συνεργασίας τῶν Πανεπιστημίων Ἀθηνῶν καὶ Κύπρου καὶ ἡ δημιουργία του χρηματοδοτήθηκε ἀπὸ τὸν Ἑπιτροπὴ Ἑρευνητικῶν Προγράμωτων τοῦ Πανεπιστημίου Κύπρου ([Παρουσίαση Προγράμματος](#)) καὶ τὸ πρόγραμμα ΠΥΒΑΓΩΡΑΣ (μὲ χρηματοδότηση τοῦ Ευρωπαϊκοῦ Κοινωνικοῦ Ταμείου καὶ Ἐθνικῶν Πρωτῶν ΕΥΡΩΔΕΚ II) ([Παρουσίαση Προγράμματος](#)). Ἡ προσπάθεια αὐτὴ χρηματοδοτήθηκε ἀπὸ τὸ πρόγραμμα «Κοινότητες» τοῦ Ἐθνικοῦ καὶ Κοινοβουλευτικοῦ Πανεπιστημίου Ἀθηνῶν (Ἔργο: «Κοινωνικὴ ἀναλλακτικὴ προσπάθεια διασφαλῆς μὲ τὸ Σώμα Ἑλληνικῶν Κειμένων»). Προγράμωτ με Κ.Α. 70/4/760. Ἑπισημητικῶς υπεσθῆναι: Διονύσιος Γούτσος.

Γιὰ νὰ δῆτε τὰς πλῆθς κητῶλογοὺς τῶν κειμένων τοῦ ΣΕΚ, πατήστε [ἄδ](#).

Γιὰ νὰ παραλάβετε ἀπὸ τὸ ΣΕΚ, ἀναφερθῆτε ἀπὸ:

ἄ. Γούτσος (2002). Σώμα Ἑλληνικῶν Κειμένων: Σχεδῶνας καὶ υλοποίηση. Πρωτῶτὸ τοῦ 6ου Διεθνοῦς Συνεδίου Ἑλληνικῆς Γλωσσολογίας, Πανεπιστήμιο Κρήτης, 18-21 Σεπτεμβρίου 2002. [Ἡλεκτρονικὸ ἔκδοτημα](#).

δ. Γούτσος (2002). The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora* 3 (1), 29-44. [Ἡλεκτρονικὸ ἔκδοτημα](#).

© 2002

Σὺμα κειμένων Ἑλληνικῶν Κειμένων

2. Phases of development



- Size: 30 million words
- Synchronic: 1990-2010
- Monolingual-non-translated texts
- Mixed: spoken-written
- Classification by text type, medium etc.
- Reference corpus of Greek: basis for analysis & comparison

2. Phases of development

- Specialized corpora
 - ▶ Greek Language Portal: http://www.greek-language.gr/greekLang/modern_greek/index.html
(newspapers, school textbooks, poems by Seferis)
 - ▶ Spoken Discourse Corpus: <http://corpus-ins.lit.auth.gr/corpus/index.html>
(everyday conversations, phone talk)
 - ▶ ongoing projects: University of Athens teaching and learner corpora

3. Major findings on Greek

- NLP applications
- quantitative/variation studies
- teaching and sociolinguistic applications

- development of new norms
- revisiting grammatical categories
- phraseology
- vocabulary change

3. Major findings on Greek

- Development of new norms: Greek connectives

1st clause position	<i>Academic</i>	<i>Opinion articles</i>	<i>Parliament speeches</i>	<i>TV interviews</i>
<i>andítheta</i> 'in contrast'	65	89	56	88
<i>ára</i> 'so'	60	56	86	97
<i>epoménos</i> 'thus'	45	52	83	88
<i>eftixós</i> 'fortunately'	66	57	50	58
<i>sinepós</i> 'consequently'	55	76	97	100
<i>entútis</i> 'however'	77	-	-	100
<i>paróla aftá</i> 'in spite'	75	-	-	75
<i>próta- próta</i> 'primarily'	-	-	67	67
<i>siberasmatiká</i> 'in conclusion'	100	100	-	-

3. Major findings on Greek

- Development of new norms: Greek connectives

2nd clause position	<i>Academic</i>	<i>Opinion articles</i>	<i>Parliament speeches</i>	<i>TV interviews</i>
<i>akrivós</i> 'precisely'	50	45	47	48
<i>áraje</i> INTER. Particle	44	77	95	100
<i>lipón</i> 'so'	80	88	87	65
<i>ómos</i> 'but'	74	88	75	60
<i>práymati</i> 'in fact'	40	38	52	45

3. Major findings on Greek

- Revisiting grammatical categories:

Fragaki (2010): Greek adjectives

- new classification:
 - classifying
 - descriptive
 - evaluative
 - deictic
 - relational etc.
- evaluative & ideological role of adjectives
- shift from one category to another:
 - e.g. περίφημος [perífimos] 'famous'
 - from positive contexts in literature
 - to negative contexts in opinion articles ('infamous')

3. Major findings on Greek

- Phraseology

Ferlas (2012): 3 to 5 word clusters in 4 Greek and English text types (CGT vs. BNC baby)

- most frequent: stance/modality clusters

ΔΕΝ ΜΠΟΡΕΪ ΝΑ: it cannot

ΔΕΝ ΠΡ'ΕΠΕΙ ΝΑ: it must not

ΘΑ ΜΠΟΡΟ΄ΥΣΕ ΝΑ: it could not

ΘΑ ΠΡ'ΕΠΕΙ ΝΑ: it should

- referential, textual, thematic etc. clusters

3. Major findings on Greek

- Phraseology

Ferlas (2012): text type preferences

e.g. literature: personal clusters

ΜΕ/ΑΠ'Ο/ΣΤΑ Μ'ΑΤΙΑ ΤΗΣ/ΤΟΥ: *with/from/in his/her eyes*

ΣΤΑ Χ'ΕΡΙΑ/Χ'ΕΡΙ ΤΗΣ/ΤΟΥ: *in his/her hand(s)*

ΚΟΥΝΗΣΕ ΤΟ ΚΕΦ'ΑΛΙ ΤΗΣ/ΤΟΥ: *moved her/his head*

ΑΠΟ ΤΟ ΣΤΟΜΑ ΤΟΥ: *from his mouth*

Η ΦΩΝ'Η ΤΟΥ/ΤΗΣ: *his/her voice*

ΓΥΡΙΣΕ ΚΑΙ ΜΕ/ΤΗΝ ΚΟ'ΙΤΑΞΕ: *turned around and looked at me/her*

3. Major findings on Greek

- Phraseology

Ferlas (2012): equivalent clusters

ΕΊΝΑΙ ΔΥΝΑΤΟΝ ΝΑ	IT IS POSSIBLE TO
ΔΕΝ ΞΕΡΩ ΑΝ	I DON T KNOW WHETHER
ΘΑ ΉΘΕΛΑ ΝΑ	I D LIKE TO
ΣΤΗΝ ΄ΑΚΡΗ ΤΟΥ	[AT/ON] THE EDGE OF THE
ΓΙΑ ΜΙΑ ΣΤΙΓΜΗ	FOR A MOMENT
ΜΕ ΤΗ ΒΟ΄ΗΘΕΙΑ [ΤΟΥ/ΤΩΝ]	WITH THE AID OF/WITH THE HELP OF

3. Major findings on Greek

- Phraseology

Ferlas (2012): non-equivalent clusters

ΚΟΥΝΗΣΕ ΤΟ ΚΕΦΑΛΙ ΤΟΥ ≠ SHOOK HIS HEAD

1	«Δεν έχεις άδικο» κούνησε το κεφάλι του συμφωνώντας. “You’re right”, he shook his head <u>in agreement</u>
2	κι αυτός κούνησε το κεφάλι του πάνω κάτω, τάχα ότι καταλάβαινε and he shook his head <u>up and down as if he understood</u>
3	Κούνησε το κεφάλι του σαν να μην το πίστευε. He shook his head <u>as if he did not believe this</u>
4	Ο Στέφανος κούνησε το κεφάλι με κατανόηση. Stephanos shook his head <u>with understanding.</u>
5	Κούνησε το κεφάλι του με αβεβαιότητα κι έσβησε το τσιγάρο He shook his head <u>with uncertainty</u> and put down his cigarette
6	και κούνησε το κεφάλι διαμαρτυρόμενος. and shook her head <u>in protest</u>

3. Major findings on Greek

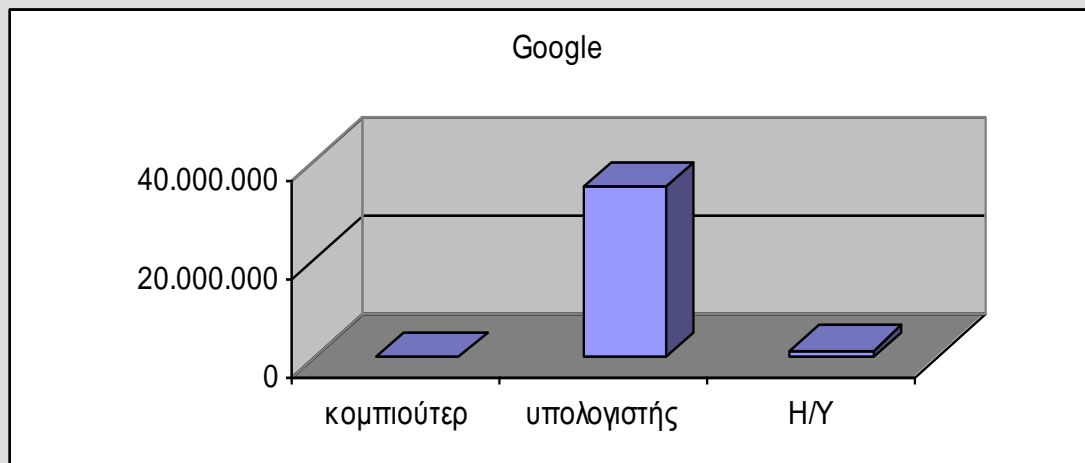
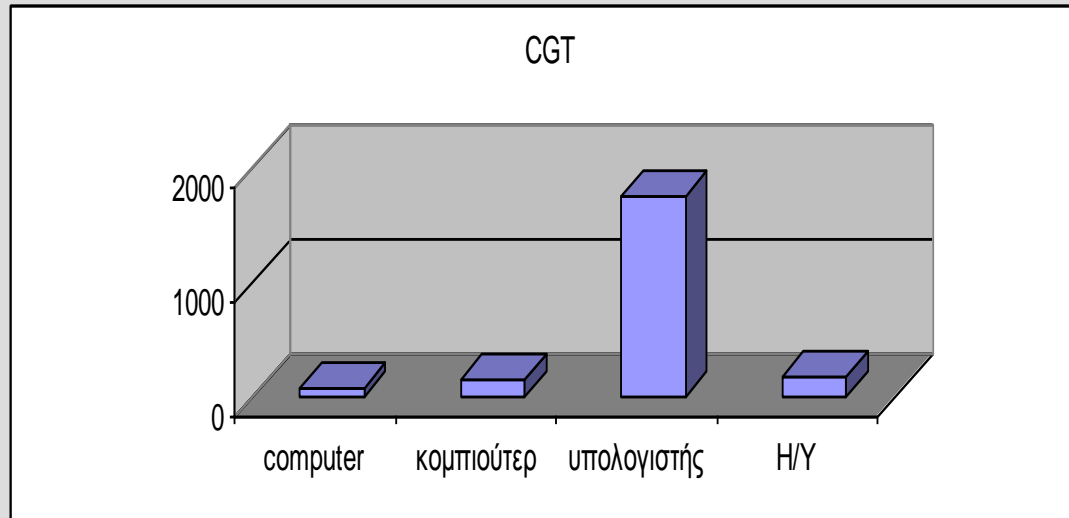
- Vocabulary change

Options for introducing new vocabulary

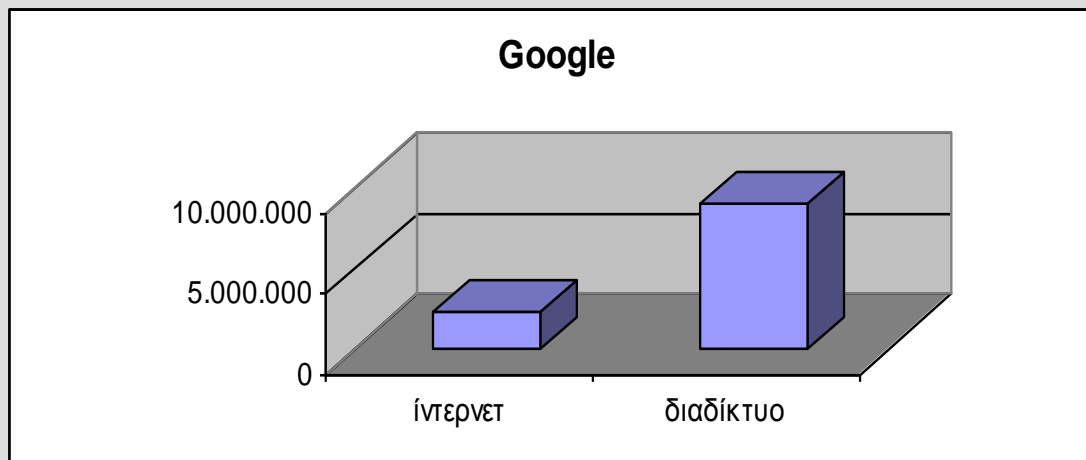
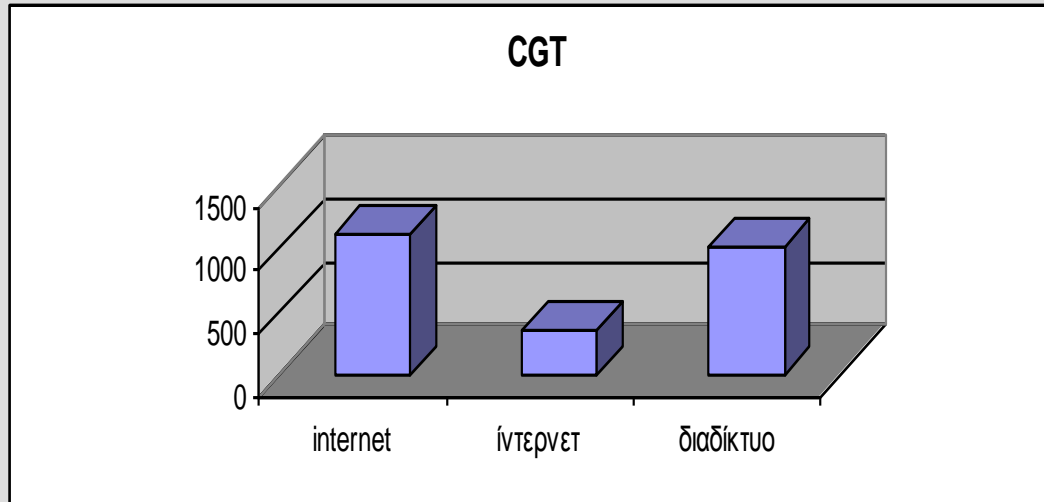
- *computer*: loan
- *κομπιούτερ* [kompjúter]: loan/transliteration
- *υπολογιστής* [ipolojistís] (= calculator): neologism

- *internet*: loan
- *ίντερνετ/ιντερνέτ* [ínternet/internét]: loan/transliteration
- *διαδίκτυο* [djadíktio] (< δια=inter, δίκτυο=network): neologism

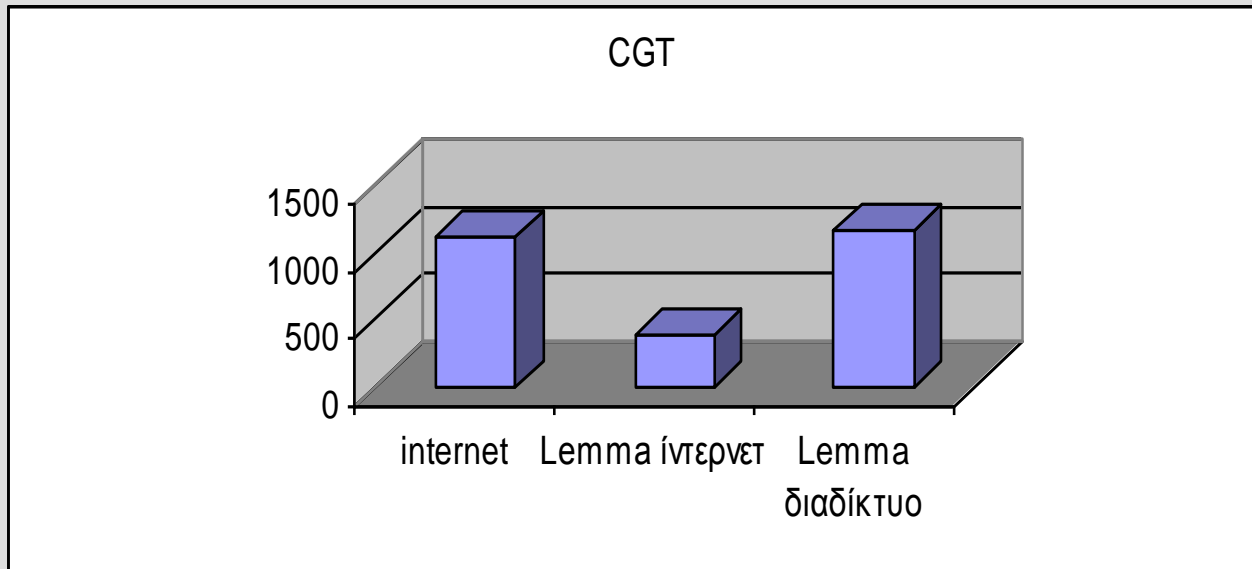
3. Major findings on Greek



3. Major findings on Greek

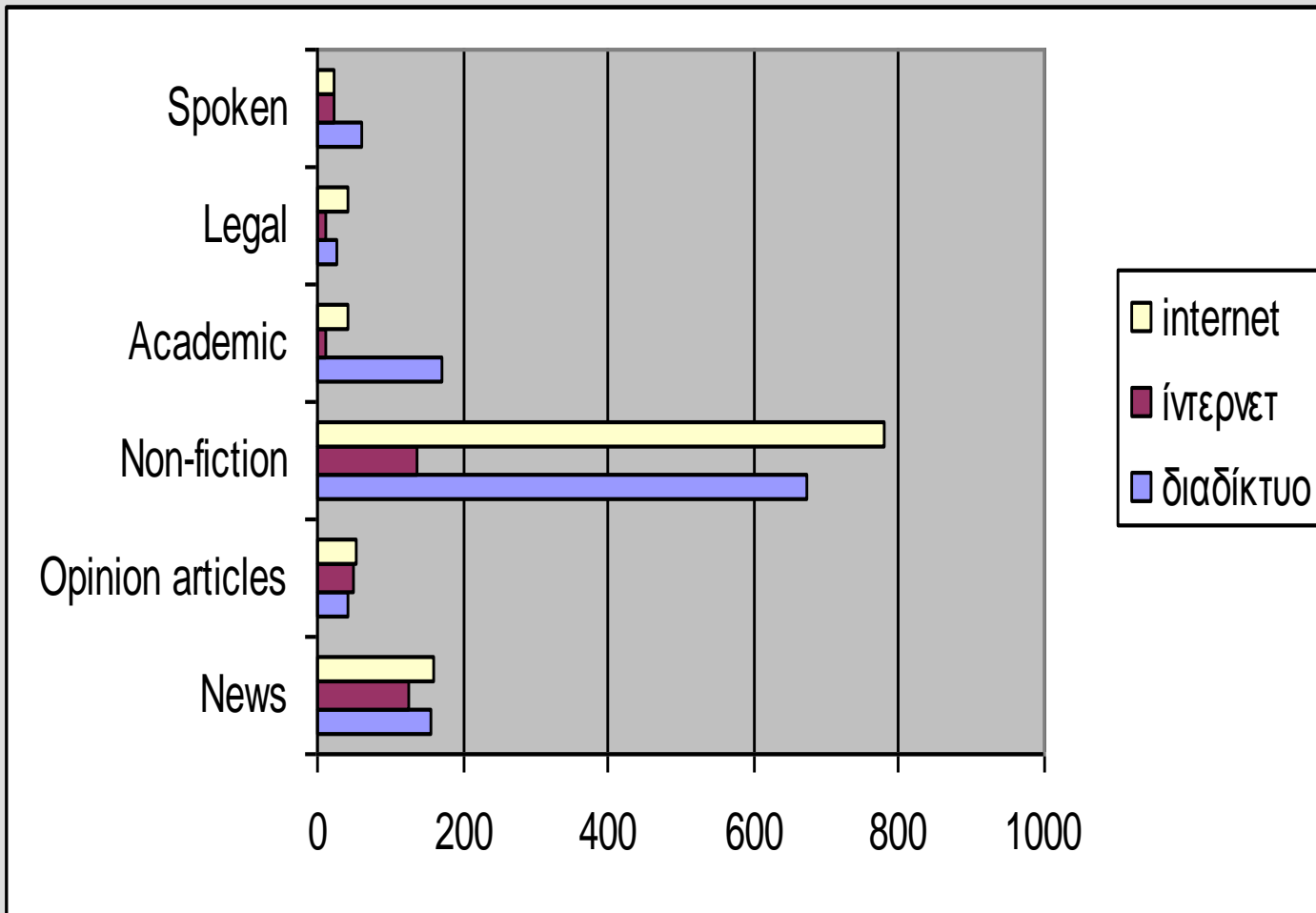


3. Major findings on Greek



ίντερνετ > ιντερνετικός, ιντερνετάκι, ιντερνέτια
διαδίκτυο > διαδικτυακός, διαδικτυωμένος, διαδικτύωση,
διαδικτυάκι

3. Major findings on Greek



4. Major needs and prospects

- Diachronic corpora
 - ▶ *Diachronic Corpus of Greek of the 20th century*:
<http://greekcorpus20.phil.uoa.gr>
 - ▶ 'polytonic' OCR
 - ▶ linking with *TLG* and Medieval Greek corpora
- Dynamic corpora: data after 2000
- NLP applications: availability, standardization

5. Greek and corpus-based translation studies

- Monolingual Greek corpora as sources of information

checking translations against Greek norms

e.g. placement of connectives

<i>ómos</i> (%)	1 st position	2 nd position	Other
Dik (1995)	21	71	8
Coelho (1999)	74	25	1
Pamouk (2007)	31	67	2

5. Greek and corpus-based translation studies

- Monolingual Greek corpora as sources of information

checking translations against Greek norms

e.g. *perífimos*

1 περιέγραφαν φαντασιώσεις που περιλαμβάνονταν στο περίφημο βιβλίο για τις σεξουαλικές παρεκκλίσεις - βιβλίο Κοέλο (1999)

2 ο Τόκατ, στον τσαϊχανέ Σενλέρ, δίπλα ακριβώς στο περίφημο λουτρό Πιφβανέ, στο μπαρ, είμαι υπεύθυνος για τ Παμούκ (2007)

3 χο στίχο, το ποίημα. Έχει γράψει τους στίχους του περίφημου ποιήματος που ξέρουμε, όταν χτυπάει η πόρτα. Παμούκ (2007)

5. Greek and corpus-based translation studies

- Monolingual Greek corpora as sources of information

checking translations against Greek norms

e.g. *κούνησε το κεφάλι του/της* 'moved his/her head'

«Όχι». Ο Μίνι κούνησε το κεφάλι του με έμφαση. (Dick 1995)

“No”. Mini shook his head no vigorously. (Dick 1991)

«Σωστά» είπε ο Κέβιν. Κούνησα το κεφάλι. (Dick 1995)

“Right”, Kevin said. I nodded. (Dick 1991)

5. Greek and corpus-based translation studies

- Monolingual Greek corpora as sources of information

comparing norms between languages

cf. Gorjanc (2006) on Slovene:

- first, only the loan word occurs in the corpus
- when the Slovene variant appears, it immediately becomes a successful rival
- the use of the loan word gradually decreases
- however: *medmrežje* has not been accepted instead of *internet*

4. Greek and corpus-based translation studies

- Bilingual corpora

From comparable corpora (Ferlas 2012) to parallel corpora (Tsoumari 2013):

e.g. frequency and range of translation choices

ST EN and

TT EL	FREQUENCY
και	1455
omitted	42
καθώς και	41
ενώ	10
αλλά και	8
ή	4
εξάλλου	1
μάλιστα	1
συνεπώς	1

ST EN but

TT EL	FREQUENCY
αλλά	16
όμως	9
ωστόσο	4
αλλά και	3
omitted	2
ενώ	1
μολονότι	1
πάντως	1

ST EN however

TT EL	FREQUENCY
ωστόσο	7
εντούτοις	2
όμως	2
μολαταύτα	1
πάντως	1

4. Greek and corpus-based translation studies

- Tsoumari (2013): development of customized annotation tools for Greek

The screenshot displays the 'Aligned Text Visual Annotation Editor' interface. The main window shows a bilingual text alignment between English (left) and Greek (right). The English text is from a 2008 press agreement, and the Greek text is its translation. The Greek text is annotated with the word 'καί' (and) in red boxes, indicating the translation of the English word 'and'. The interface includes a menu bar (File, Window), a toolbar with navigation and annotation tools, and a right-hand panel with 'TT ADDITION' and 'CONTEXT' sections. The 'TT ADDITION' section contains fields for 'TT EL', 'TT EL Expression', 'TT Rhetorical Relation', 'TT Category', 'TT Phrase-level Connection', 'TT Position', 'TT Rendering of', 'TT Analysis / Rendering of Text/Expression', 'ST Clue', 'Additional TT EL', 'TT Comments', and 'ST Comments'. The 'CONTEXT' section contains fields for 'ST Verb', 'ST Adjective', 'ST Adverb', 'ST Other', 'ST More', 'ST Less', 'TT Verb', 'TT Adjective', 'TT Adverb', 'TT Other', 'TT More', and 'TT Less'. The status bar at the bottom shows the collection path and document name.

4. Greek and corpus-based translation studies

- Conclusions:

more work is needed in the analysis of Greek monolingual corpora

> identification of norms

> comparison with translated text norms (check for simplification, explicitation, normalization, levelling-out; cf. Baker 1996)

more and more varied bilingual corpora, involving Greek and a broad range of other languages, are needed

corpus building and corpus analysis are closely intertwined