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

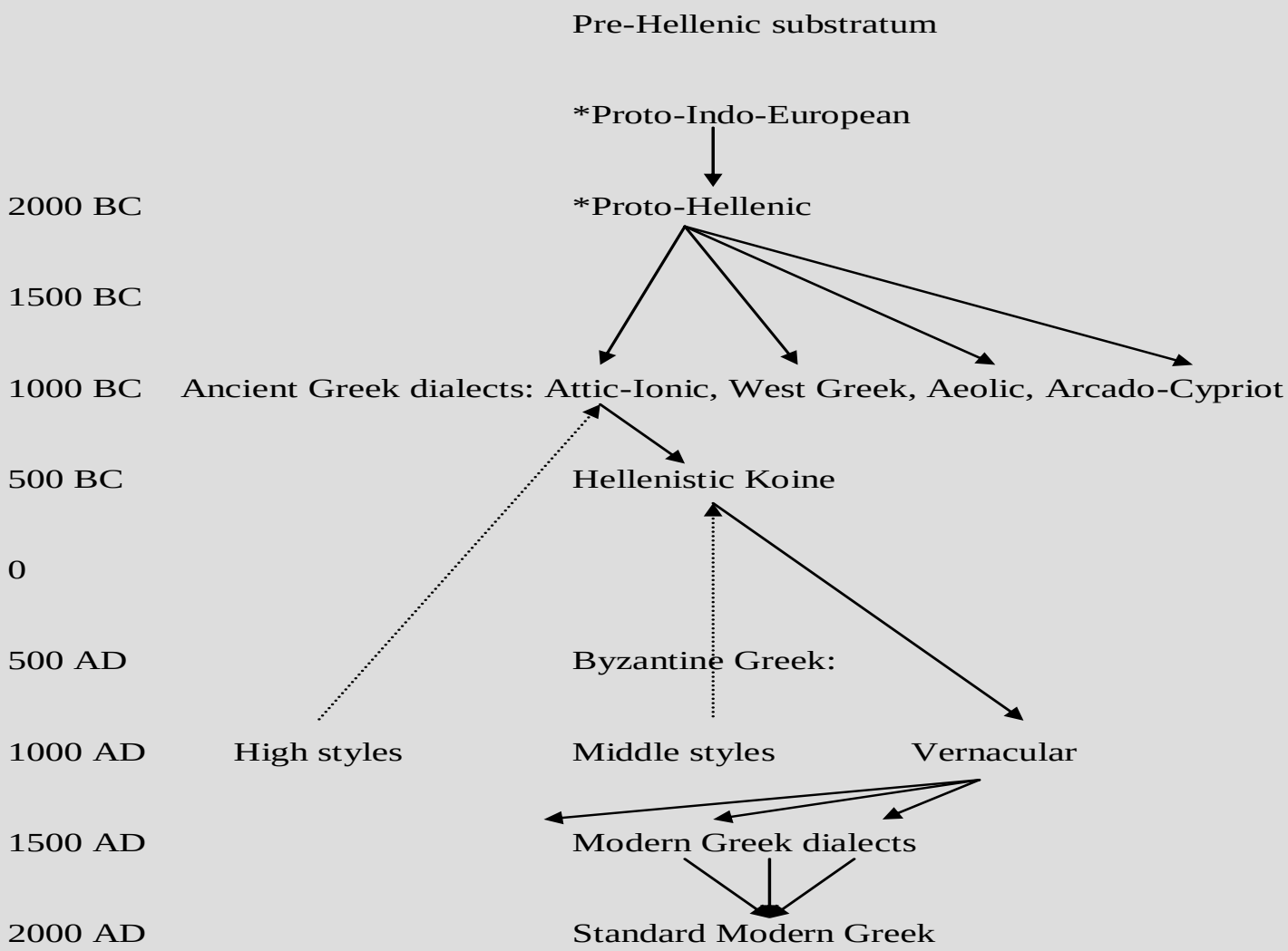
- *Corpus-based analysis of Greek*

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 co-ordination 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

Greek corpora

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

Γ159	"Νένα μου, *λέγει ἡ Ἀρετή*, φρόνιμα δασκαλεύγεις,
Γ171	"Παιδί μου, *λέγει ἡ νένα της*, σφάνουσι τὰ λογιάζεις,
Γ299	"Νένα μου, *λέγει ἡ Ἀρετή*, ἴντά 'ν' τὰ δασκαλεύγεις;
Ε613	"Νένα, *τσῆ λέγει ἡ Ἀρετή*, τὸ γίνηκεν ἐγίνη
Ε1133	"Ἄμε πέ, *λέγει ἡ Ἀρετή*, γλήγορα τοῦ κυροῦ μου

0908	λύσει,/ καὶ σὰν ὄντας κρατεῖς	κερὶ κι ἄνεμος σοῦ τὸ σβήσει.// ἐκεῖνος ὁποὺ
1145	[οὐδὲ καπνίσματα μποροῦν οὐδὲ	κεριὰ οὐδὲ θρόνοι,// τὸν κόσμο ὅλο
0415	νὰ μ' ἀφήσει·/ ἐπά 'ν τ' ἀφτούμενο	κερὶ ποὺ μελετᾶς νὰ σβήσεις,// καὶ τὸ κορμὶ ὁποὺ
0408	μηδ' εἶδα το ποτέ μου,/ **μά 'ναν**	**κερὶν ἀφτούμενο ἐκράτουν κι ἤσβησέ μου.//**

(a) *Sacrifice of Abraham*

Α0756 ΠΟΙ	κ' ἐπόνει/ σὰν τὸ	κερὶ ἀνελίγωνε κ' ἐφύρα σὰν τὸ χιόνι.//
Ε0923 ΡΩΤ	ἀγάλια ἐχάνετο σὰν τὸ	κερὶ ὄντε σβήνη,/ ἤκλαψα κ' ἐλυπήθηκα πολλὰ
Γ1102 ΦΡΟ	κι ὡσὰ φυσήξης τὸ	κερὶ ὁπού 'ψες, νὰ τὸ σβήσης// καὶ δῆ το καὶ
Β2391 ΠΟΙ	κι οὐδὲ ψυχάρι οὐδὲ	κερὶ οὐδὲ φωτιὰ οὐδὲ γράμμα/ τοῦ ἐπόμεινε
Β0757 ΠΟΙ	κι αὐτὸς στὴν κεφαλὴν ἕνα	κερὶ σβημένο·/ τὸν Ἄνεμον ἀνάδια του ἤδειχνε
Ε0414 ΠΟΙ	τῶν ἀμματιῶν του ἡ λάμψη/	κερὶ σβηστὸν ἐφύλαγε καὶ πάγει ἐδὰ νὰ τ' ἄψη.//
Β2177 ΠΟΙ	οἱ τρεῖς τως σ' μιὰ μερά,	Κερί, Ψυχάρι, Ἀμάξι,/ μὲ πεθυμιὰ ἀνιμένασιν
Γ1396 ΡΩΤ	δὲ σ' εἶδα ποτέ μου,/ **μὰ ἕνα**	**κερὶν ἀφτούμενον ἐκράτουν κ' ἤσβησέ μου.//**
Δ1763 ΠΟΙ	λαβώση·// σὰ νά 'χεν εἶσται	κέρινο, τέτοιας λογῆς διαβαίνει,/ στὸν κάμπο

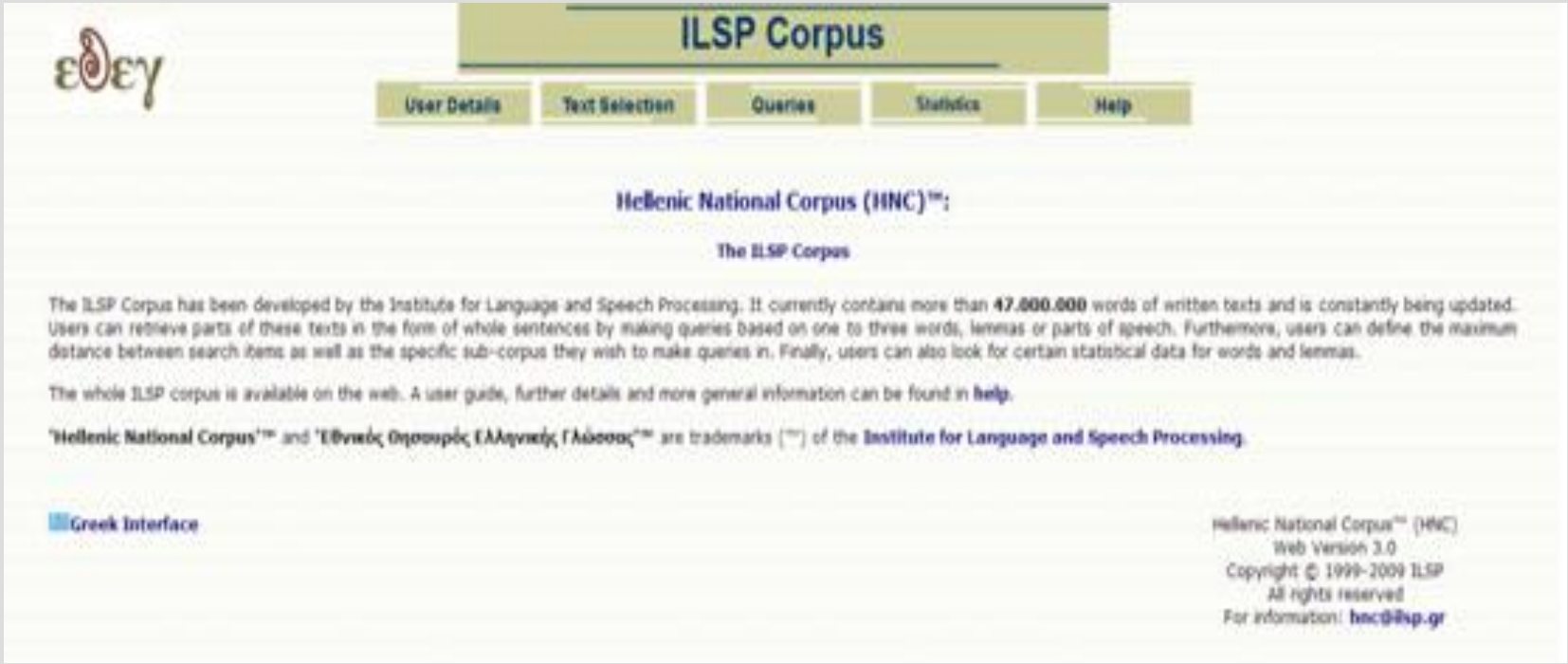
(b) *Erotókritos*

2. Phases of development

Greek corpora

- 1990s-2000s: two ‘super-corpora’

  Hellenic National Corpus: http://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

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	*Academic*	*Opinion articles*	*Parliament speeches*	*TV interviews*
andíθeta 'in contrast'	65	89	56	88
ára 'so'	60	56	86	97
epoménos 'thus'	45	52	83	88
eftiχós 'fortunately'	66	57	50	58
sinepós 'consequently'	55	76	97	100
entútis 'however'	77	-	-	100
paróla aftá 'in spite'	75	-	-	75
próta- próta 'primarily'	-	-	67	67
siberasmatiká 'in conclusion'	100	100	-	-

3. Major findings on Greek

- Development of new norms: Greek connectives

2nd clause position	*Academic*	*Opinion articles*	*Parliament speeches*	*TV interviews*
akrivós 'precisely'	50	45	47	48
áraje INTER. Particle	44	77	95	100
lipón 'so'	80	88	87	65
ómos 'but'	74	88	75	60
práymati '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 ΜΠΟΡΕΊ ΝΑ: it cannot
    - ΔΕΝ ΠΡΈΠΕΙ ΝΑ: it must not
    - ΘΑ ΜΠΟΡΟΎΣΕ ΝΑ: it could not
    - ΘΑ ΠΡΈΠΕΙ ΝΑ: it should

  - referential, textual, thematic etc. clusters

- Phraseology

Ferlas (2012): text type preferences

e.g. literature: personal clusters

ΜΕ/ΑΠΌ/ΣΤΑ ΜΆΤΙΑ ΤΗΣ/ΤΟΥ: *with/from/in his/her eyes*
ΣΤΑ Χ'ΕΡΙΑ/Χ'ΕΡΙ ΤΗΣ/ΤΟΥ: *in his/her hand(s)*
ΚΟΥΝΗΣΕ ΤΟ ΚΕΦΆΛΙ ΤΗΣ/ΤΟΥ: *moved her/his head*
ΑΠΟ ΤΟ ΣΤΌΜΑ ΤΟΥ: *from his mouth*
Η ΦΩΝΉ ΤΟΥ/ΤΗΣ: *his/her voice*
ΓΎΡΙΣΕ ΚΑΙ ΜΕ/ΤΗΝ ΚΟΊΤΑΞΕ: *turned around and looked at me/her*

- Phraseology

Ferlas (2012): equivalent clusters

Ε'ΙΝΑΙ ΔΥΝΑΤ'ΟΝ ΝΑ	IT IS POSSIBLE TO
ΔΕΝ Ξ'ΕΡΩ ΑΝ	I DON T KNOW WHETHER
ΘΑ 'ΗΘΕΛΑ 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

- 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 in agreement
2	κι αυτός **κούνησε το κεφάλι** του πάνω κάτω, τάχα ότι καταλάβαινε and he shook his head up and down as if he understood
3	**Κούνησε το κεφάλι** του σαν να μην το πίστευε. He shook his head as if he did not believe this
4	Ο Στέφανος **κούνησε το κεφάλι** με κατανόηση. Stephanos shook his head with understanding.
5	**Κούνησε το κεφάλι** του με αβεβαιότητα κι έσβησε το τσιγάρο He shook his head with uncertainty and put down his cigarette
6	και **κούνησε το κεφάλι** διαμαρτυρόμενος. and shook her head in protest

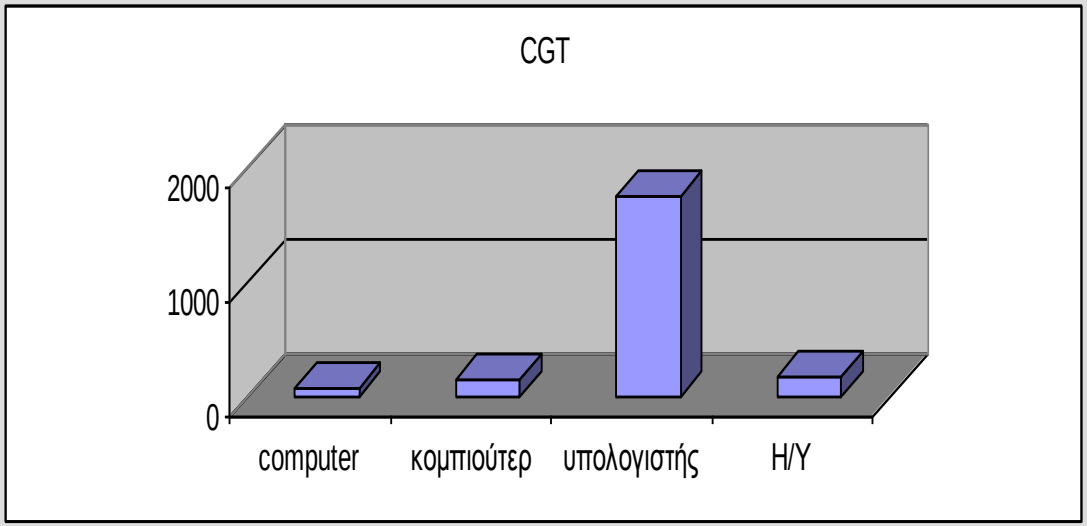
- Vocabulary change

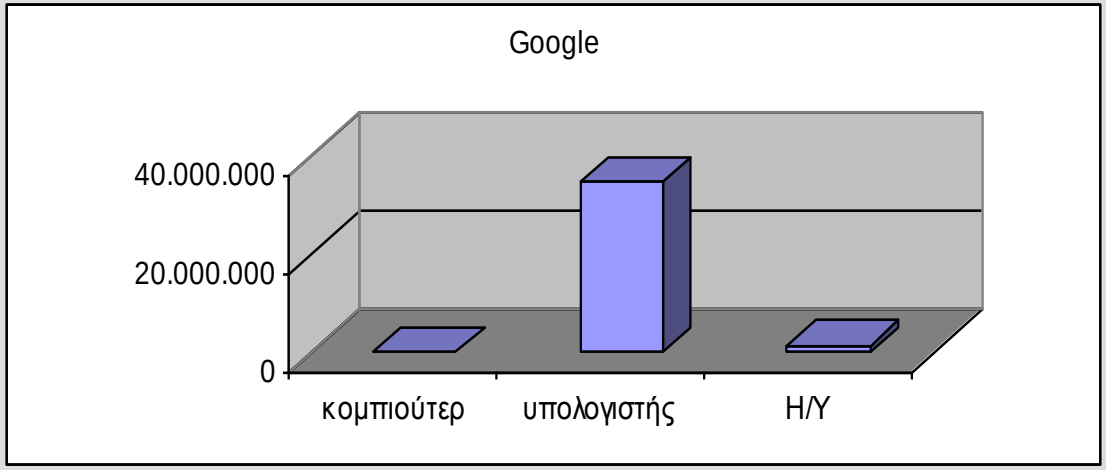
Options for introducing new vocabulary

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

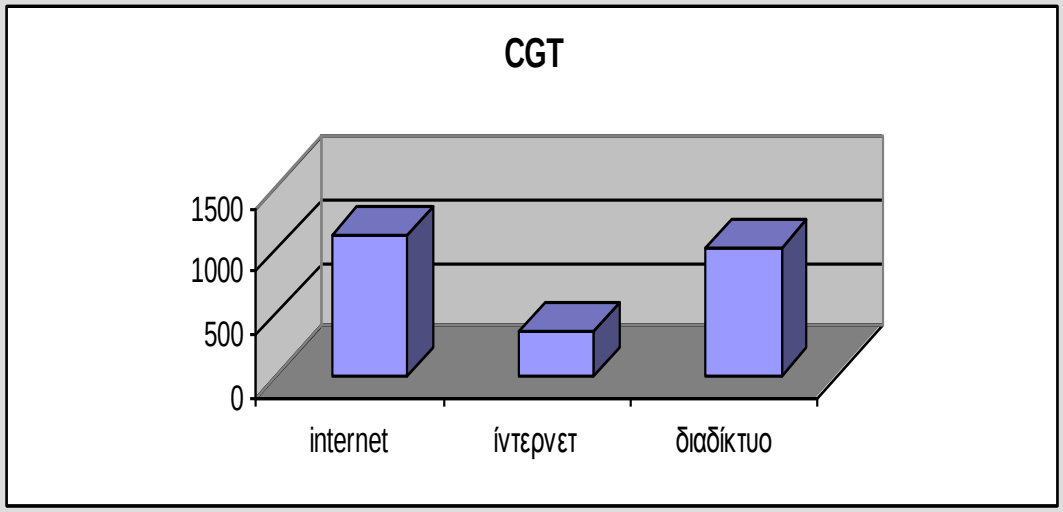
- *internet*: loan
- *ίντερνετ/ιντερνέτ* [índernet/internét]: loan/transliteration
- *διαδίκτυο* [ðjaðíktio] (< δια=inter, δίκτυο=network): neologism

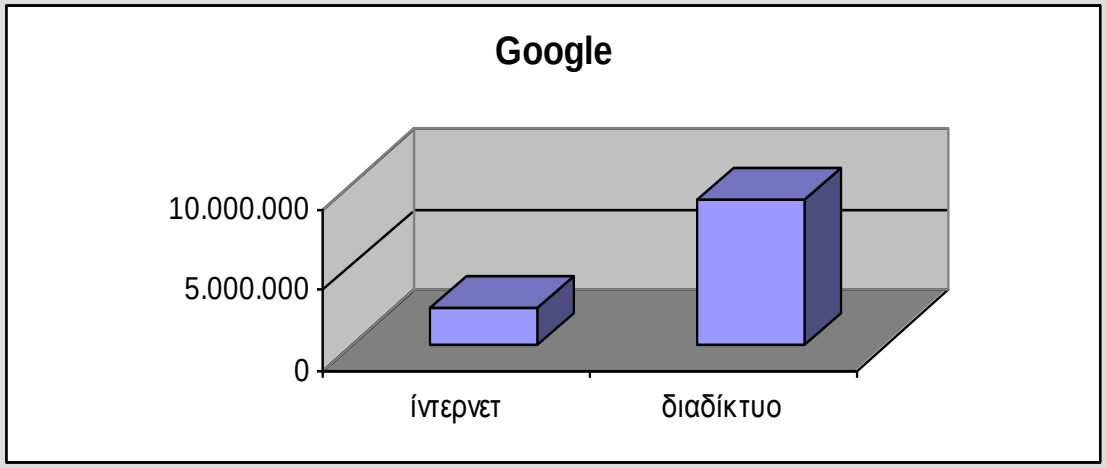
3. Major findings on Greek



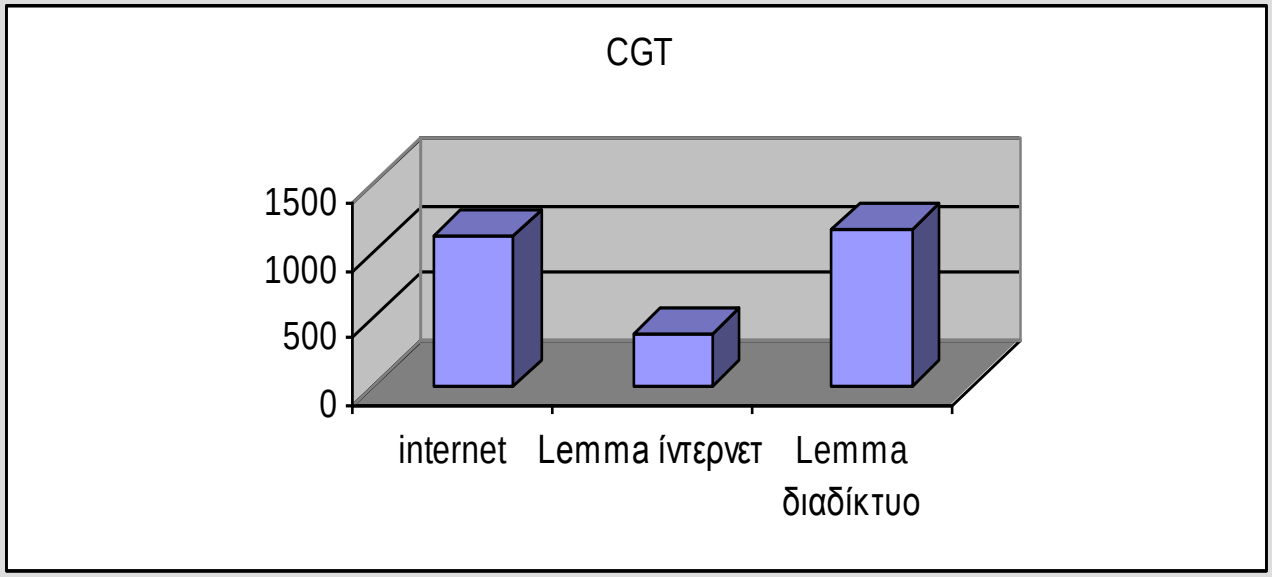


3. Major findings on Greek





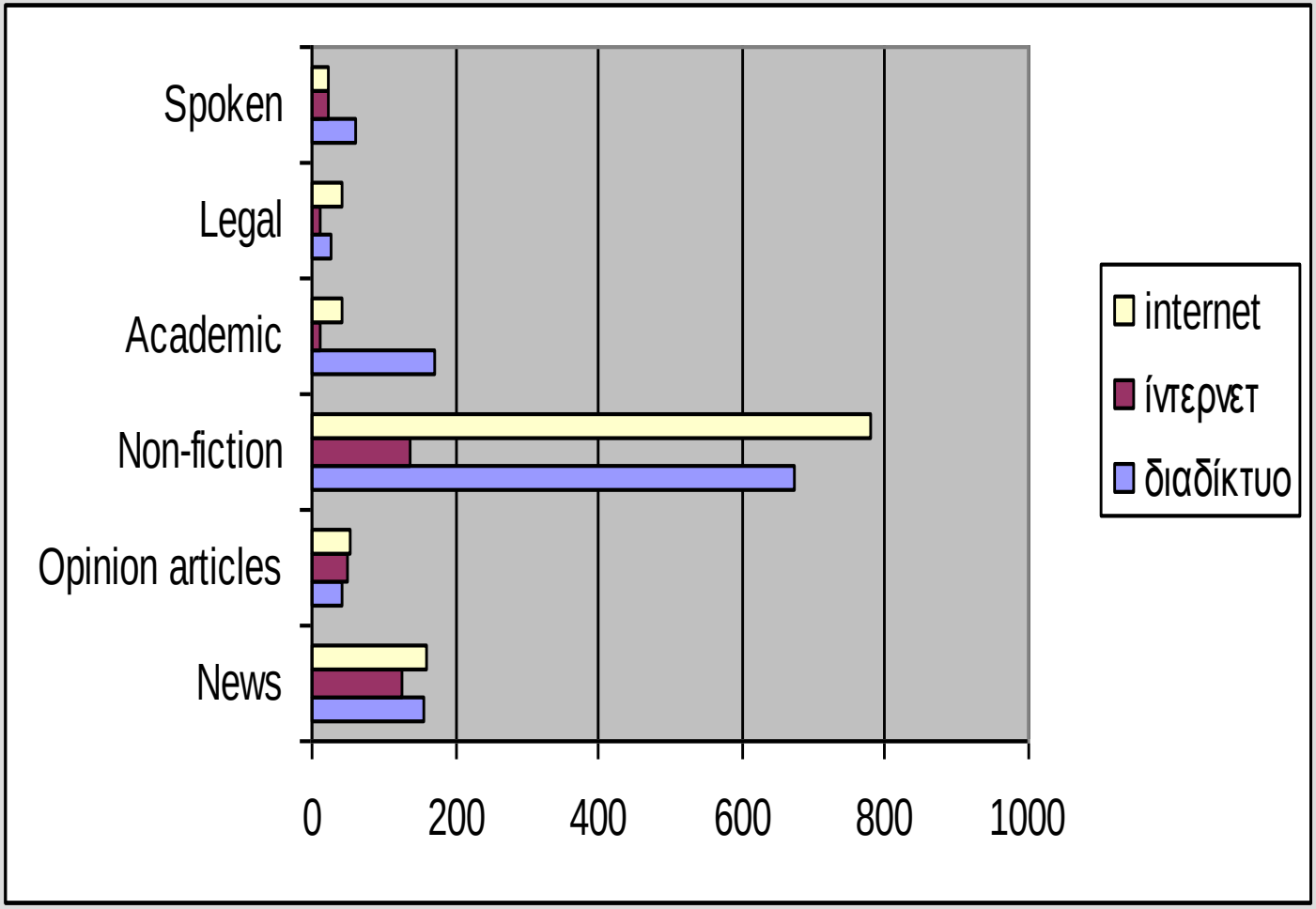
3. Major findings on Greek



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

ómos *(%)*	1st position	2nd 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**

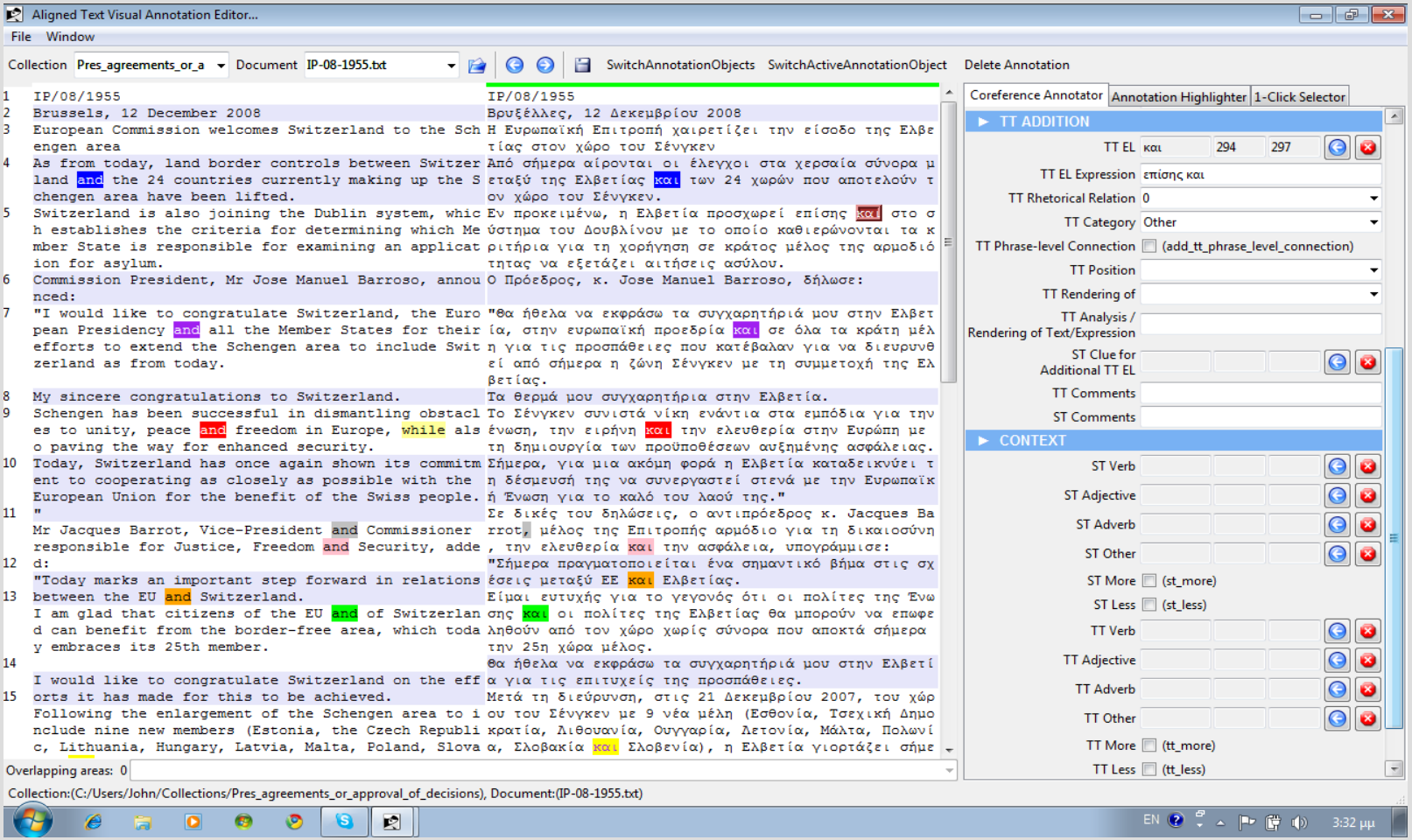
TT EL	FREQUENCY
αλλά	16
όμως	9
ωστόσο	4
αλλά και	3
omitted	2
ενώ	1
μολονότι	1
πάντως	1

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



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