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

Pre-Hellenic substratum

*Proto-Indo-European

*Proto-Hellenic

2000 BC

1500 BC

1000 BC  Ancient Greek dialects: Attic-Ionic, West Greek, Aeolic, Arcado-Cypriot

500 BC  Hellenistic Koine

0

500 AD  Byzantine Greek:

1000 AD  High styles  Middle styles  Vernacular

1500 AD  Modern Greek dialects

2000 AD  Standard Modern 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
2. Phases of development

Greek corpora

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

- 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**

<table>
<thead>
<tr>
<th>1st clause position</th>
<th>Academic</th>
<th>Opinion articles</th>
<th>Parliament speeches</th>
<th>TV interviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>andíðeta ‘in contrast’</td>
<td>65</td>
<td>89</td>
<td>56</td>
<td>88</td>
</tr>
<tr>
<td>ára ‘so’</td>
<td>60</td>
<td>56</td>
<td>86</td>
<td>97</td>
</tr>
<tr>
<td>epoménos ‘thus’</td>
<td>45</td>
<td>52</td>
<td>83</td>
<td>88</td>
</tr>
<tr>
<td>eftixós ‘fortunately’</td>
<td>66</td>
<td>57</td>
<td>50</td>
<td>58</td>
</tr>
<tr>
<td>sinepós ‘consequently’</td>
<td>55</td>
<td>76</td>
<td>97</td>
<td>100</td>
</tr>
<tr>
<td>entútis ‘however’</td>
<td>77</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>paróla aftá ‘in spite’</td>
<td>75</td>
<td>-</td>
<td>-</td>
<td>75</td>
</tr>
<tr>
<td>próta- próta ‘primarily’</td>
<td>-</td>
<td>-</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>siberasmatiká ‘in conclusion’</td>
<td>100</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### 3. Major findings on Greek

- Development of new norms: Greek connectives

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

15
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 ΜΠΟΡΕΙ ΝΑ: it cannot
  ΔΕΝ ΠΡΕΠΕΙ ΝΑ: it must not
  ΘΑ ΜΠΟΡΟΥΣΕ ΝΑ: it could not
  ΘΑ ΠΡΕΠΕΙ 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/ΑΠΌ/ΣΤΑ ΜΆΤΙΑ ΤΗΣ/ΤΟΥ: with/from/in his/her eyes
ΣΤΑ ΧΈΡΙΑ/ΧΈΡΙ ΤΗΣ/ΤΟΥ: in his/her hand(s)
ΚΟΥΝΗΣΕ ΤΟ ΚΕΦΆΛΙ ΤΗΣ/ΤΟΥ: moved her/his head
ΑΠΌ 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
3. Major findings on Greek

- Phraseology

Ferlas (2012): equivalent clusters

ΕΊΝΑΙ ΔΥΝΑΤΌΝ ΝΑ  IT IS POSSIBLE TO
ΔΕΝ ΞΈΡΩ ΑΝ  I DON T KNOW WHETHER
ΘΑ ΉΘΕΛΑ ΝΑ  I D LIKE TO
ΣΤΗΝ ΆΚΡΗ ΤΟΥ  [AT/ON] THE EDGE OF THE
ΓΙΑ ΜΙΑ ΣΤΙΓΜΗ  FOR A MOMENT
ΜΕ ΤΗ ΒΟΉΘΕΙΑ [ΤΟΥ/ΤΩΝ]  WITH THE AID OF/WITH THE HELP OF
3. Major findings on Greek

- Phraseology

Ferlas (2012): non-equivalent clusters

| 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 |
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* [djádiktio] (< δια=inter, δίκτυο=network): neologism
3. Major findings on Greek
3. Major findings on Greek
3. Major findings on Greek

internet> интернетикός, интернетάκι, интернетέτια
диαδίκτυο> διαδικτυακός, διαδικτυωμένος, διαδικτύωση,
διαδικτυάκι
3. Major findings on Greek
4. Major needs and prospects

- Diachronic corpora
  - *Diachronic Corpus of Greek of the 20\textsuperscript{th} 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

<table>
<thead>
<tr>
<th>ómos (%)</th>
<th>1st position</th>
<th>2nd position</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dik (1995)</td>
<td>21</td>
<td>71</td>
<td>8</td>
</tr>
<tr>
<td>Coelho (1999)</td>
<td>74</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>Pamouk (2007)</td>
<td>31</td>
<td>67</td>
<td>2</td>
</tr>
</tbody>
</table>
5. Greek and corpus-based translation studies

- Monolingual Greek corpora as sources of information
  checking translations against Greek norms

  e.g. *perífimos*

  1. periégrafan fantasiwseis pou peri lambánontan sto períphimo biblio gia tis seξουαλικές parékklíseis - biblio Koélo (1999)
  2. o Tókat, ston tsáïxané Senvléρ, dípla akribwís sto períphimo loutró Pívbavé, sto μπαρ, eïmaí upéðúnoς gia t Pamoúk (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

<table>
<thead>
<tr>
<th>ST EN <strong>and</strong></th>
<th><strong>FREQUENCY</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>και</td>
<td>1455</td>
</tr>
<tr>
<td>omitted</td>
<td>42</td>
</tr>
<tr>
<td>καθώς και</td>
<td>41</td>
</tr>
<tr>
<td>ενώ</td>
<td>10</td>
</tr>
<tr>
<td>αλλά και</td>
<td>8</td>
</tr>
<tr>
<td>ή</td>
<td>4</td>
</tr>
<tr>
<td>εξάλλου</td>
<td>1</td>
</tr>
<tr>
<td>μάλιστα</td>
<td>1</td>
</tr>
<tr>
<td>συνεπώς</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ST EN <strong>but</strong></th>
<th><strong>FREQUENCY</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>αλλά</td>
<td>16</td>
</tr>
<tr>
<td>όμως</td>
<td>9</td>
</tr>
<tr>
<td>ωστόσο</td>
<td>4</td>
</tr>
<tr>
<td>αλλά και</td>
<td>3</td>
</tr>
<tr>
<td>omitted</td>
<td>2</td>
</tr>
<tr>
<td>ενώ</td>
<td>1</td>
</tr>
<tr>
<td>μολονότι</td>
<td>1</td>
</tr>
<tr>
<td>πάντως</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ST EN however</th>
<th><strong>FREQUENCY</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>ωστόσο</td>
<td>7</td>
</tr>
<tr>
<td>εντούτοις</td>
<td>2</td>
</tr>
<tr>
<td>όμως</td>
<td>2</td>
</tr>
<tr>
<td>μολαταύτα</td>
<td>1</td>
</tr>
<tr>
<td>πάντως</td>
<td>1</td>
</tr>
</tbody>
</table>
4. Greek and corpus-based translation studies

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