Treebanking in the World of Thucydides

Linguistic annotation for the Hellespont Project

Francesco Mambrini

Center For Hellenic Studies

Deutsches Archäologisches Institut

November 20 2012

Outline

1. What digital corpora for Ancient History?
   - The questions at hand
   - Data-driven approaches

2. Linguistic Annotation of Thucydides 1.98-118
   - The Hellespont Project
   - Examples

Outline

1. What digital corpora for Ancient History?
   - The questions at hand
   - Data-driven approaches
2. Linguistic Annotation of Thucydides 1.98-118
   - The Hellespont Project
   - Examples

A web of knowledge

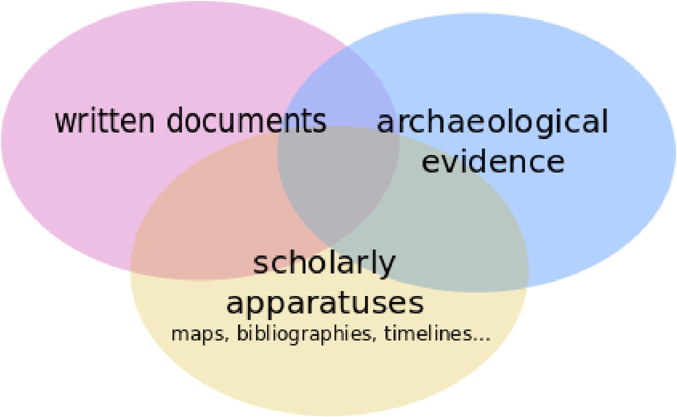


Figure: A simplified model

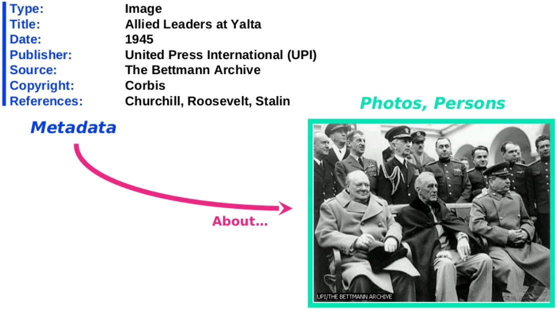
Interconnectedness: the problem

The multivalent nature of historical thought [. . .] eludes the keyword-indexed approach to the Web today on offer through Google and other search engines. Though we can summon up an exhaustive list of Web resources that contain the words "Gallipoli" and "sources", today's Web cannot effectively respond to a basic historical question such as, "which sources attest the Gallipoli Campaign of World War I?"

B. Robertson

CIDOC Conceptual Reference Model

Objects represented as being part of events



Type:	Image
Title:	Allied Leaders at Yalta
Date:	1945
Publisher:	United Press International (UPI)
Source:	The Bettmann Archive
Copyright:	Corbis
References:	Churchill, Roosevelt, Stalin

Figure: by Doer and Stead 2009

One more problem!

Know what our sources are!

- big and complex works; e.g. Thucydides:
  - 6.126 sentences, 167.512 words
  - ca 30 years of war, + 50 years in digression, references that go back to before the Trojan War!
- Unstructured natural language
- Written in Ancient Greek
- Controversial (interpretation and textual reconstruction)
- Literary work (= shaped by discursive and ideological strategies)

Outline

1. What digital corpora for Ancient History?
   - The questions at hand
   - Data-driven approaches
2. Linguistic Annotation of Thucydides 1.98-118
   - The Hellespont Project
   - Examples

Ontologiemodellierung für die Erforschung von Ritualstrukturen (SBF 619, Heidelberg)

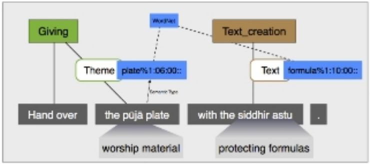


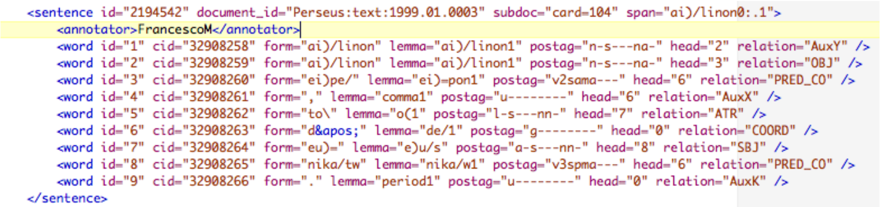
Figure: Event extraction from texts

NLP Pipeline

NLP Process	Ancient Greek?
Chunking	🙂
Lemmatization	😐
POS-tagging	😐
Syntactic parsing	😐
Word-sense disambiguation	🙁
Co-reference resolution	🙁
Semantic role annotation	🙁

Using and Enhancing the available resources

The Ancient Greek Dependency Treebank



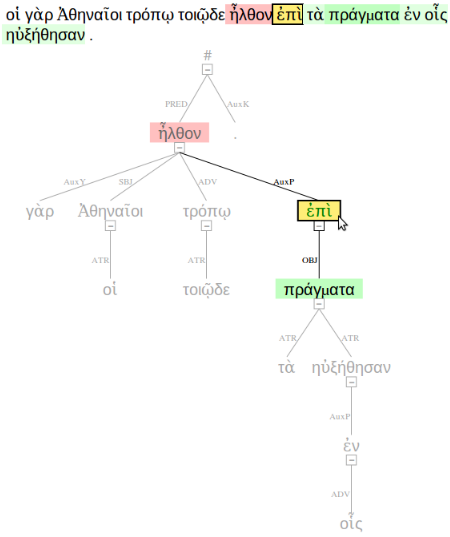
```
<sentence id="2194542" document_id="Perseus:text:1999.01.0003" subdoc="card=104" span="ai)/linon0:.1">
    <annotator>FrancescoM</annotator>
    <word id="1" cid="32908258" form="ai)/linon" lemma="ai)/linon1" postag="n-s---na-" head="2" relation="AuxY" />
    <word id="2" cid="32908259" form="ai)/linon" lemma="ai)/linon1" postag="n-s---na-" head="3" relation="OBJ" />
    <word id="3" cid="32908260" form="ei)pe/" lemma="ei)=pon1" postag="v2sama---" head="6" relation="PRED_CO" />
    <word id="4" cid="32908261" form="," lemma="comma1" postag="u--------" head="6" relation="AuxX" />
    <word id="5" cid="32908262" form="to\" lemma="o(1" postag="l-s---nn-" head="7" relation="ATR" />
    <word id="6" cid="32908263" form="d&apos;" lemma="de/1" postag="g--------" head="0" relation="COORD" />
    <word id="7" cid="32908264" form="eu)=" lemma="e)u/s" postag="a-s---nn-" head="8" relation="SBJ" />
    <word id="8" cid="32908265" form="nika/tw" lemma="nika/w1" postag="v3spma---" head="6" relation="PRED_CO" />
    <word id="9" cid="32908266" form="." lemma="period1" postag="u--------" head="0" relation="AuxK" />
</sentence>
```

AGDT: treebank with word-by-word morphological and dependency-based syntactical description

a step forward: semantic information

A syntactic tree

Thuc. 1.89.1



οἱ γὰρ Ἀθηναῖοι τρόπῳ τοιῷδε ἦλθον ἐπὶ τὰ πράγματα ἐν οἷς ηὐξήθησαν .

Outline

1. What digital corpora for Ancient History?
   - The questions at hand
   - Data-driven approaches
2. Linguistic Annotation of Thucydides 1.98-118
   - The Hellespont Project
   - Examples

A case study
Athens, 479-431 BCE

Goal:

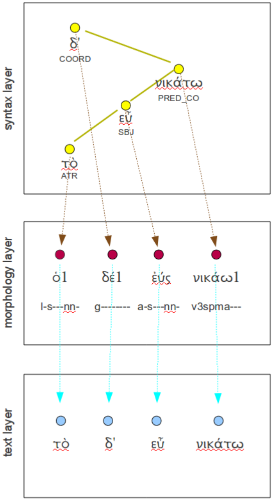
- Connecting textual and archaeological sources in the Perseus DL and Arachne via CIDOC-CRM

Steps:

- Enriching the text of one source (Thucydides) with linguistic and historical information
- Identify and mark events on the text
  - manually
  - data-driven approach
- Integrating secondary literature (through data mining algorithms)

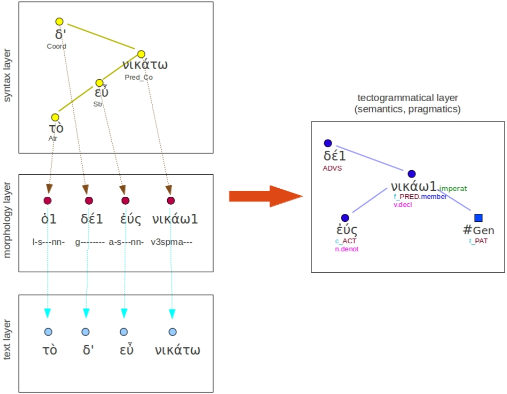
Toward a 3-level scenario

Morphology and Syntax



Toward a 3-level scenario

+ semantic and pragmatical information



Outline

1. What digital corpora for Ancient History?
   - The questions at hand
   - Data-driven approaches

2. Linguistic Annotation of Thucydides 1.98-118
   - The Hellespont Project
   - Examples

With tectogrammatical annotation:

Our text is:

1. easier to browse for content-related search (easier to use in digital environments)
2. more informative on historically relevant questions

With tectogrammatical annotation:

Our text is:

1. easier to browse for content-related search (easier to use in digital environments)
2. more informative on historically relevant questions

With tectogrammatical annotation:

Our text is:

1. easier to browse for content-related search (easier to use in digital environments)
2. more informative on historically relevant questions

Conclusions

1. Currently, our literary sources are not structured for semantic, event-based queries
2. NLP processes for event extraction are not yet capable of handling raw Ancient Greek texts
3. NLP tools and techniques are adaptable to the task
    - provide standards
    - help and speed manual annotation
    - (incidentally) they add a lot of information on linguistic aspects of the documentary sources