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

written documents
archaeological evidence
scholarly apparatuses
maps, bibliographies, timelines...

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
Objects represented as being part of events

**Metadata**

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

**Photos, Persons**

**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
<table>
<thead>
<tr>
<th>NLP Process</th>
<th>Ancient Greek?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chunking</td>
<td>😊</td>
</tr>
<tr>
<td>Lemmatization</td>
<td>😞</td>
</tr>
<tr>
<td>POS-tagging</td>
<td>😞</td>
</tr>
<tr>
<td>Syntactic parsing</td>
<td>😞</td>
</tr>
<tr>
<td>Word-sense disambiguation</td>
<td>😞</td>
</tr>
<tr>
<td>Co-reference resolution</td>
<td>😞</td>
</tr>
<tr>
<td>Semantic role annotation</td>
<td>😞</td>
</tr>
</tbody>
</table>
Using and Enhancing the available resources

The Ancient Greek Dependency Treebank

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

**a step forward:** semantic information
What digital corpora for Ancient History?
Linguistic Annotation of Thucydides 1.98-118

The questions at hand
Data-driven approaches

A syntactic tree
Thuc. 1.89.1
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 pragmatically information

What digital corpora for Ancient History?
Linguistic Annotation of Thucydides 1.98-118
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
Currently, our literary sources are not structured for semantic, event-based queries.

NLP processes for event extraction are not yet capable of handling raw Ancient Greek texts.

NLP tools and techniques are adaptable to the task, providing standards, helping and speeding manual annotation, and (incidentally) adding a lot of information on linguistic aspects of the documentary sources.