DHQ: Digital Humanities Quarterly

2016
Volume 10 Number 2

Treebanking in the world of Thucydides. Linguistic annotation for the Hellespont Project

Francesco Mambrini <fmambrini_at_dainst_dot_de>, Deutsches Archäologisches Institut, Berlin

Abstract

The Hellespont project (DAI, Tufts University) aims to structure the text of a passage from the ancient Greek historian Thucydides (1.89-118), in order to highlight events, persons and peoples that populate the world of the author and connect the different digital sources available for their study. Event annotation in the text in particular requires an in-depth linguistic analysis of morphology, syntax and semantics. However, the available resources for Ancient Greek do not provide adequate standards to support the encoding of semantic and pragmatic phenomena in Ancient Greek texts. In this paper, we discuss the motivation of the project and how we adapted the so called *tectogrammatical annotation* of the Prague Dependency Treebank to identify the events and describe their structure. The linguistic notion of *valency*, which is central to tectogrammatical sentence representation, proves very useful for this analysis of Ancient Greek.

As the papers presented at the 2013 Digital Classics Association conference "Words, Space, Time" and collected in this volume attest, digital resources for the study of the Greek and Roman civilizations are rapidly growing in quantity and quality. In many fields of Classical Studies (such as digital archaeology or epigraphy), progress has been spectacular. In other areas, however, the gap between the resources that can be applied for studying the ancient world and the state-of-the-art technologies for processing and exploring data is still considerable. In terms of Natural Language Processing (NLP), for instance, the resources for classical languages are still far less numerous and sophisticated than those available for modern languages such as English. 1

In the case of Ancient Greek and Latin, the creation of linguistically annotated corpora, which has been carried out by several institutions in the last decade, has brought at least some of the tools and resources that are commonly used in modern computational linguistics to the attention of the Classicists. In current linguistic research, corpora that integrate a fine-grained annotation of morphology and syntax, as well as potentially many other levels of analysis, are known as *treebanks* [Abeillé 2003]. Started between 2007 and 2008, projects like the Ancient Greek and Latin Dependency Treebank (Perseus Project at Tufts University), the PROIEL treebank [Haug and Jøhndal 2008], the Index Thomisticus Treebank (Catholic University, Milan) of Mediaval Latin [Passarotti 2009] have already reached quite a mature status. All of the mentioned projects provide several Greek or Latin texts with complete annotation on the morphology of each word and on the syntactic structures of each sentence; all the annotated texts are encoded in machine-readable format and freely downloadable from the websites of the different projects[1]. However, syntax and morphology are the only level that are annotated in the Greek and Latin treebanks that are currently published: other areas of linguistic studies, such as semantics or pragmatics are still not represented. Moreover, while the available treebanks provide a set of schemas and guidelines for the treatment of morpho-syntacitc phenomena, no standards or manuals for the annotation of semantics and pragmatics of the ancient languages exist that can serve as a model to extend the current schemas. 2

This is the situation that we found when, in the context of a project entitled *Hellespont* (see Section 1), we decided to perform an in-depth linguistic annotation of a section of an ancient Greek literary text. The goal of this project was twofold. On the one hand, we wished to produce a comprehensive linguistic analysis in a digital format that would enable readers to get a better understanding of the historical events that are narrated in the text. On the other hand, we intended to lay the foundations for a complex linguistic description of Ancient Greek 3

that could also take semantic and pragmatic phenomena into account, along with syntax and morphology.

In what follows, we will present the motivations (Sections 1.1 and 1.2), as well as the methodology, models and solutions that we have adopted (Section 2). 4

1 The Hellespont Project

1.1 Introduction: views, entities and events

The *Hellespont Project* (http://hellespont.dainst.org/) was launched in 2009 by the Deutsches Archäologisches Institute (DAI) and the Perseus Project at Tufts University. The team, led by A. Thomas (DAI and University of Cologne), built on the experiments performed by Thomas and others at the Perseus Project on linking textual and archaeological sources, as well as modern scholarly works, in a digital edition of Thucydides [Romanello and Thomas 2012]. 5

The project targeted a limited period and context, the history of Athens in the years between the end of the Persian Wars (479 BCE) and the outburst of the Peloponnesian War (431 BCE). One single textual source of the events was selected, that of the Athenian historian Thucydides, who deals with the period in a digression (chapters 89 to 118) of Book One of his *Histories*; since Antiquity, this digression has been known as the *Pentecontaetia* (Greek for “fifty-year period”)[2]. In terms of raw numbers, the segment is composed of 178 sentences, for a total of 4641 words[3]. 6

Hellespont aimed to answer two fundamental questions. First, how should a literary text be structured, so that the relevant data can be accessed, browsed and retrieved by historians? Second, how should digitized documents be presented so that they leverage the power of a digital environment and show the meaningful connections between different sources? 7

The first task was carried out as a part of a bigger effort to link the data from the main collections of the two sponsoring institutions, the archaelogical database Arachne (http://arachne.uni-koeln.de/) and the Perseus Digital Library (http://www.perseus.tufts.edu).[4] 8

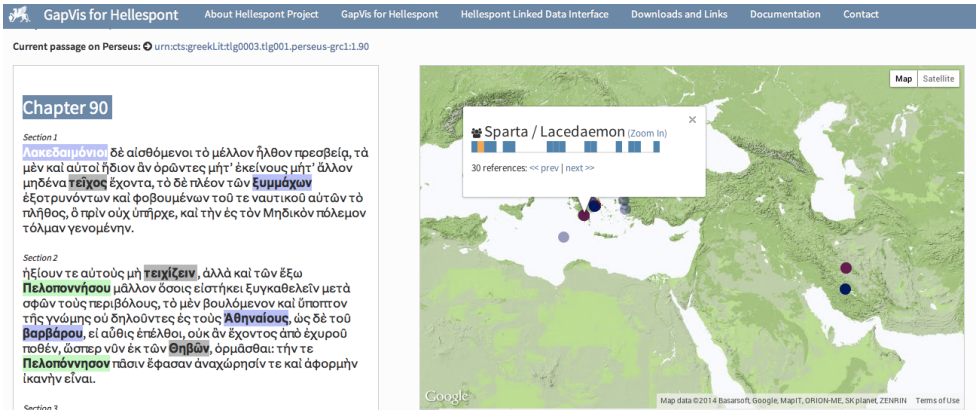

Figure 1. Hellespont, reading view

The most important part of this task was the annotation of the *named entities* mentioned in the ancient source. The names of places, persons, and peoples were annotated and disambiguated, and they were then linked to authority files and gazetteers; this process of disambiguation allowed us to display the text of Thucydides and all the other data in our collections, such as monuments or other objects in Arachne, that include references to the same entities in a unified reading environment. The interface that was chosen for visualizing the structured text of Thucydides and to highlight the connections between it and the materials in Arachne and Perseus is based on the model of GapVis (http://nrabinowitz.github.io/gapvis). 9

The work of Thucydides is browsable in multiple different views. In the *Reading* view, the passages of the text are displayed with the named entities highlighted: different colors are used to distinguish between the classes of entities (persons, geographical names, man-made landmarks, organizations); on a side pane, those entities that are linked to a geographic location (e.g. the Spartan people, which are annotated as a collective organization, 10

but are clearly identifiable with the city of Sparta) are represented on a map (Figure 1). The *Entity detail* view lists all the linked data from Arachne, the iDAI gazetteer (https://gazetteer.dainst.org/), the Perseus Project and other external repositories, such as Pleiades (http://pleiades.stoa.org/); the most important co-refenced entities in Thucydides' *Pentecontaetia* are also listed for readers and visualized on a map (Figure 2).

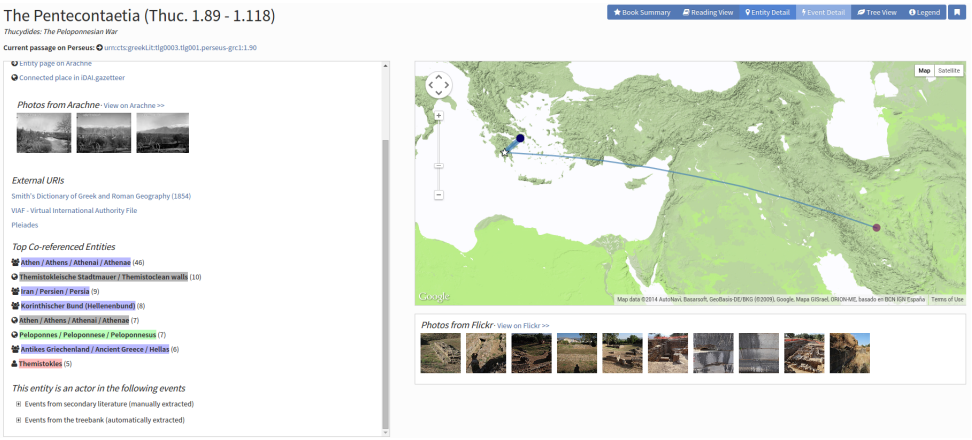

Figure 2. Hellespont, entity detail view

1.2 Events and the event view

Entities in history are not important per se except as they are involved in a series of human activities. 11

At first glance, it would seem obvious that a work of historical prose dedicated to the war between the two superpowers of its times should mostly contain narrations of events and that consequently it should not be difficult to recognize and identify them in the flow of the text. The reference model that *Arachne* is currently adopting to link the different objects in its collections, the event-based CIDOC-CRM (http://www.cidoc-crm.org/), seems also ideal to leverage this notion and bridge the world of texts and archaeological *realia*. CIDOC-CRM is precisely built on the idea that the common involvement in the same historical "event" or activity is a point of contact between monuments, objects, actors, and places [Doerr 2003]. 12

The designers of Hellespont decided to proceed in the same direction and to thus also include passages of ancient texts within the data that are linked in the model. For example, if we know that one object (the walls of Athens) and one actor (e.g. Themistocles) are involved in the same event (the Athenians, following Themistocles' advice, rebuilt the city fortifications destroyed by the Persians), and if we identify a specific passage in a textual source (Thucydides, *Histories*, 1.89.3) that makes reference to that episode, the annotation would provide an effective link between the archaeological traces of a given monument and the text excerpt. Then, by setting up a chronology of the annotated events (either absolute or relative), it becomes possible to create a timeline that allows readers to access the related sources in the collection. 13

Thucydides' text has been read as a source for the history of the Peloponnesian War for centuries. Therefore, several readers in the history of scholarship have already undertaken the task of indexing and summarizing the text according to the most relevant content of each paragraph. These summaries, that are found in the printed editions and translations of Thucydides, can be adopted as a guide for event annotation. The portion of the text between *Histories* 1.102.1 and 1.102.3, for example, is often paraphrased in the indexes of Thucydides as: "the Athenians summoned to Ithome, then dismissed"; thus, this label can be used to isolate a part of the text (the two paragraphs) as referring to one event (the siege of Ithome), or two events (the Athenians intervention in the siege; their dismissal). 14

This approach was adopted for a first stage of Hellespont; it was then implemented in the *event view* within the GapVis interface, where the content of the *Pentecontaetia* can be accessed from a timeline; each event is further described in its structural relations with other related episodes (which is particularly useful for skirmishes or operations that are part of a larger campaign) and in its internal articulation (its actors, or its temporal and geographical frame)[5]. The historical commentary of Hornblower (1991) was chosen as the main reference for the identification of the events. 15

Different scholars may differ considerably even in their segmentation of a text, and the choice for one solution over the others might be arbitrary. Yet the main problem with this approach is often one of granularity. In some sections of the text, especially in the second half after chapter 98, the style of the narration becomes quick and rather unadorned [Hornblower 1991, 149]: many facts appear to be piled up one after the other. In such cases, readers are forced to summarize chapters and paragraphs using general labels, while the stages and the articulation of the facts remain difficult to capture. Thus, for example chapter 1.105 and 1.106 are summarized by Hornblower (1991, 164-6) as "operations in the First Peloponnesian War": in the context of event annotation, a reader would preferably know the exact nature of these operations, rather than be left with such an unspecific reference. 16

In such cases, the reader that wants to isolate the stages of an event with greater accuracy is forced to look closely at the linguistic structure of each sentence. Consequently, instead of taking textual units like chapters or paragraphs as a basis for the annotation, we were forced to consider the linguistic articulation of the text itself. This task of isolating real-world events in natural language, however, is extremely challenging. An example from Thucydides (taken from the aforementioned narration of the Athenian intervention in Ithome) may help in illustrating the kind of pitfalls that we encountered: 17

> *μάλιστα δ' αὐτοὺς ἐπεκαλέσαντο ὅτι τειχομαχεῖν ἐδόκουν δυνατοὶ εἶναι, τοῖς δὲ πολιορκίας μακρᾶς καθεστηκυίας τούτου ἐνδεᾶ ἐφαίνετο: βίᾳ γὰρ ἂν εἷλον τὸ χωρίον*
>
> [The siege of Ithome proved tedious, and the Lacedaemonians called in, among other allies, the Athenians…] [They] invited them especially because [they] considered [them] particularly skilled in siege operations, while, since the siege for them was dragging on, [their] own deficiency in that sort of warfare was clear: for otherwise [they] would have taken the place by force

If we consider the two sentences in isolation, we may start from the plausible assumption that the main verbs (*ἐπεκαλέσαντο*: called in; *ἐφαίνετο*: was clear; *εἷλον*: would have taken) should lead us to identify a core set of the narrated events. Yet it is immediately clear that the last one of them does not belong to the domain of history, but the verb mood characterizes the event as the unrealistic outcome of a counterfactual hypothesis, both in Greek and in the English translation. On the other hand, even the first verb, which may seem to those familiar with Greek history to point to the main fact that is referred to here (the Athenians are summoned to the siege of Ithome), is not unproblematic. If we consider the sentence that precedes, which is reported in brackets at the beginning of the translation, we see that Thucydides is not reporting a new event, he's instead alluding for the second time to the same episode he narrated in the previous sentence; the main focus of the first quoted sentence is the cause that brought about the Athenian intervention, while the fact that they were called in is simply repeated from the preceding statement. This distinction, which may seem trivial to the human reader and is most likely performed unconsciously by them, requires a subltler linguistic distinction between given and new information. 18

The internal analysis of the kind that we would like to implement for our event view (with the actors involved and temporal and geographical frames) is also more complex than it may seem at first. In the given English translation, all the actors involved, except from the Lacedaemonians, are mentioned only by pronouns ("they" or "them"). Their identity cannot be guessed without considering the whole context[6]. In the Greek original, however, it is not even a question of resolving the reference of the pronouns, as it is in English: the bracketed "theys" of the translation are entirely omitted in the Greek original, as it is often the case with verb and noun arguments (and syntactic subjects in particular) that can be identified in the context. 19

Therefore, if we want to assess how many events we can identify in these sentences and what the named entities involved in them are, we must consider a rather large spectrum of linguistic phenomena, from part-of-speech tagging, to verbal modality, co-refence resolution, and syntactic structure. Syntactic analysis, however, is not sufficient: subjects of active verbs and agents of passive verbs are obviously distinguished in syntax, and yet these arguments play the same role in the performance of an action, as in the two sentences: "Caesar was killed by Brutus" and "Brutus killed Caesar". A definition of the *semantic role* played by each argument must then be introduced in addition to the labels for syntactic function. Annotation of information structure can also be used to distinguish between *new* and *given* information (as in the case of the repeated "they were called" of our example). 20

Our task, thus, became that of providing a comprehensive linguistic annotation of all these phenomena, so that our annotated corpus may support an alternative, more data-driven analysis than the one based on summaries found in the scholarly literature. In other words, we intended to create a treebank that could be effectively used to answer questions such as: "What are all the actions that the Athenians undertake in Thucydides' *Pentecontaetia* (regardless whether the Athenians are explicitly mentioned by name in the text or just implied from the context)?,""Is there a difference between what they do and what the Spartans do?""Who is the most influential person in the *Pentecontaetia*?""Is the role of individuals more or less important that that of the collective actors?" 21

Although a thorough discussion of these questions would be out of the scope of our paper, we believe that the multi-layer treebank that we have created support the kind of analysis that we were aiming at (see in particular Section 3 and figure 4). 22

2 A Multi-layer Treebank

2.1 The Ancient Greek Dependency Treebank

As we have already said, treebanks are digital corpora that embed word-by-word annotation on linguistic phenomena, including at least part-of-speech tagging and syntax. For classical studies, a number of treebanks of Greek and Latin (listed above) have been recently made available to the public. Among them, the *Ancient Greek and Latin Dependency Treebank (AGLDT)* promoted by the Perseus Project aims in principle to include all the extant literature in the two ancient languages. 23

Currently, the Greek section (*AGDT*, v.1.7) contains almost 375,000 annotated words, including the complete text of the Homeric poems, the extant opera of Hesiod and Aeschylus, as well as five tragedies of Sophocles and the *Euthyphro* of Plato. All the texts are chunked into sentences and a complete morphological analysis (including lemmatization) of each word is provided [Bamman et al. 2009]. The syntactic function of the words in a sentence is described according to a formalism inspired by dependency grammar [Tesnière 1959]; in this representation (which may be contrasted with a constituency-based formalism, where words are grouped into phrases), the words are put directly in relation with each other; thus, a subject is made to depend directly on its governing verb, articles and attributes on the nouns they are referred to, and so on. The set of relations between heads and dependents can thus be visualized as a dependency tree, with acyclic, directed edges connecting the words (see Figure 3 below). The annotation of morphology and syntax is provided for every word in each text, including for example particles and conjunctions, as well as for punctuation marks[7]. The morphological annotation and the type of dependency grammar adopted by Perseus are modeled on two layers (the morphological and the so-called *analytical layer*) of the *Prague Dependency Treebank* of Czech (PDT)[8]. 24

At present, sentence- and word-tokenization of the Ancient Greek texts included in the Perseus Digital Library is fully automated, nonetheless no NLP-tool is available that provides reliable automatic tagging and syntactic parsing[9]; therefore the work of annotation is entirely manual. In spite of these limitations, the AGDT provides an excellent standard that supports the morphological and syntactic analysis of Greek. 25

Taking advantage of this model has turned out to be the best practice for our own work. Our task was thus twofold. On the one hand, we used the format and the rules of the AGDT to annotate the syntax and morphology of the *Pentecontaetia*. On the other hand, we still needed to look for other possible models to supplement the analysis with the kind of semantic/pragmatic information (such as semantic roles, valency, co-reference resolution and information structure) that the format of the AGDT does not support. 26

2.2 Tectogrammatical Sentence Representation

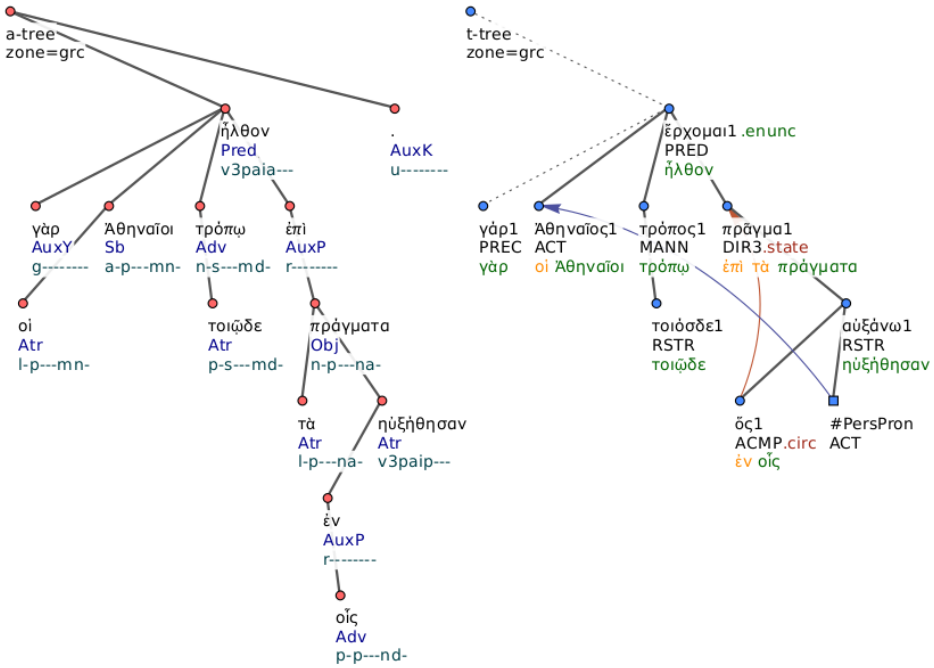

Figure 3. Analytical (left) and tectogrammatical (right) tree of Thucydides, *Histories*, 1.89.1

As we have seen, the treatment of syntax in the AGDT follows the model of analytical annotation of surface syntax used by the *Prague Dependency Treebank*. The PDT, which is built upon a theoretical framework known as *Functional Generative Description* [Sgall et al. 1986], implements also a third level of annotation that is aimed at capturing the meaning of a sentence. This layer, called *tectogrammatical*, encodes many of those linguistic phenomena that we hoped to describe in the text of Thucydides [Sgall et al. 2004]. 27

When we started considering this model for our work, tectogrammatical annotation had been implemented only for Czech, Arabic and partially for English [Cinková et al. 2008]. Its adaptation to Greek proved to be a complex process, both in the definition of a set of guidelines for the different grammatical constructions and in the concrete work of textual annotation, which cannot be discussed in full here[10]. In what follows, we will limit ourselves to describing some basic concepts of tectogrammatical sentence representation. 28

The overall meaning of the sentence is captured by the set of head-dependent relations and by a combination of other factors. In particular, a tectogrammatical tree still represents the structure of a sentence as a dependency tree, which is not different from what happens with the standard syntactic annotation of the AGDT (compare the two trees of Figure 3). In the prototypical cases, the main verb dominates all the other words, both in analytic and tectogrammatical trees. The nodes of a tectogrammatical tree, however, do not represent all the words that are attested in the text, but only those which carry an independent lexical meaning. Functional words, like prepositions and conjunctions, or those words that introduce special nuances of meaning, like modal verbs, auxiliaries, or articles, are not reproduced as nodes, but rather as properties of the lexical nodes[11]. 29

The head-dependent relations are described by a set of labels (the so-called *functors*), that aim to describe the semantic relation between the words. Thus, instead of syntactic roles, such as *(direct or indirect) object* or *adverbial*, the functors describe the complements as *actor*, *patient*, *addressee* or *benefactor*[12]. The co-reference of the linguistic "pointers" (such as personal or demonstrative pronouns) is also annotated: each pronoun is linked directly to the words it refers to. 30

In tectogrammatical representation, words that are left out because they can be recovered from the context are integrated with the introduction of artificial nodes. The most important type of reconstructed nodes are the *valency arguments*, and, most typically in Greek, the implied subjects. In tectogrammatical annotation, the concept of *valency* does indeed play a crucial role. The valency of a word may be described as the list of the obligatory arguments that are required in order to fill each of the particular meanings of a given word; Tesnière 31

(1959: 102) speaks of the valency arguments as the actors that play a part in the "drama" evoked by a word, and distinguish them from the *circumstantials* that are used to set the temporal, spatial or discursive frame of an event. Valency arguments can be omitted whenever they can be easily implied, but are semantically always required; they are therefore always integrated in the tectogrammatical trees. To come back to the example from Thucydides quoted above (1.102.2, see Section 1.2), all the "they" and "them" within brackets in the English translation would correspond to newly inserted nodes in the tectogrammatical representation of the Greek original, for both subjects and (direct) objects belong to the valency of the Greek verbs used in the sentences.

32 One crucial feature in tectogrammatical annotation is that even the co-reference of the reconstructed nodes is marked: in the example of Thucydides, *Histories*, 1.102.2, tectogrammatical annotation enables us to link not only the pronouns that are explicitly used in the Greek original with either the Spartans or the Athenians, but also the implied subjects of the verbs. Much more than a modern translation, therefore, tectogrammatical sentence representation can serve to disambiguate the interpretation of this passage.

33 Figure 3 (tree on the right) provides a clear example of a newly reconstructed node with co-reference resolution in a tectogrammatical tree. The subject of the relative clause *ἐν οἷς ηὐξήθησαν* ("in which [they] prospered") is missing in Greek. However, since it is a required argument in the valency of the verb *αὐξάνω* (to prosper), an artificial node is created, with the artificial lemma *#PersPron* that is generally used for personal pronouns; in the visualization of Figure 3, this node is recognizable by its square shape, which distinguishes it visually from the round nodes of the words attested in the text. Furthermore, this artificial node is linked to the word that it refers to, as it can be seen by the arrow that connects it to the subject of the main clause (*οἱ Ἀθηναῖοι*, the Athenians): the implied "they" of this relative clause are none other than the Athenians.

34 The sets of arguments required by (the different meanings of) each verb in Thucydides, *Histories*, 1.89-118 were registered in a special *valency lexicon*, essentially a dictionary that lists all the arguments that are requested to fill each of the senses of a word [13]. Conversely, each verb in the treebank was annotated with the ID of the valency frame that is appropriate to the specific meaning of the word. The valency frames in the dictionary are also linked each to the specific meaning, and therefore to the specific English translation, of the corresponding lemma in the *Greek English Lexicon* of Liddell and Scott [LSJ 1940], digitized in the Perseus Digital Library [Rydberg-Cox 2002]. In this way, every verb that is attested in our passage and is assigned a valency frame is also effectively connected to the appropriate meaning and translation in an authoritative dictionary.

3 Conclusions

35 Tectogrammatical annotation is a complex task. The representation of a sentence at this level and the interplay with different surface phenomena is language-specific: the rules that have been identified for the tectogrammatical sentence representation of Czech and (partially) English are not immediately applicable to Ancient Greek without thorough investigation. The process of annotation and, especially, the writing of a set of guidelines for annotation are still in progress. Readers that are interested can download the latest version of our treebank (and valency lexicon) and check the current status of annotation (http://hellespont.dainst.org/startpage/downloads.html).

36 Nonetheless, the annotation that we have performed, especially on the valency frame of verbs and co-reference resolution, already allows for a complex analysis of the linguistic "dramas" [Tesnière 1959, 102] in Thucydides, *Histories*, 1.89-118. By using our annotations, for example, we can already extract all the verbs where the Athenians fulfill the semantic role of *actors*; this search is not limited to those contexts where the Athenians are explicitly mentioned, but can be easily extended to include the passages where pronouns are used or even when the Athenians are the implied arguments. Moreover, by using the definitions of the LSJ lexicon that is linked with the valency frame in our dictionary, we can choose to present the results of our investigations in English translation, even if the data relate to the content of the original Greek text. A visualization of the results for such a query, i.e. the English translations of all the verbs that have the Athenians or a co-referenced node as the semantic agent in the *Pentecontaetia*, is given in the word-cloud of Figure 4.

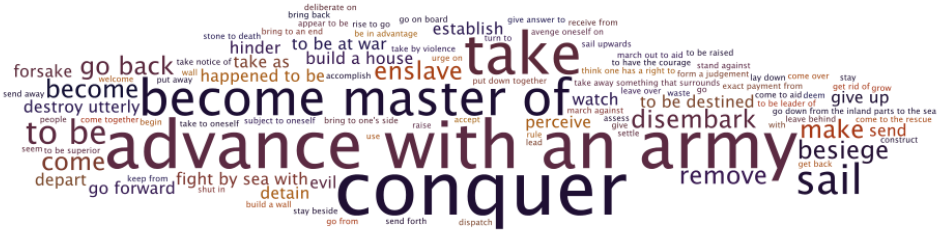

Figure 4. Verbs of the *Pentecontaetia* that have the Athenians as semantic agents; the translations are taken from the LSJ dictionary

Tectogrammatical annotation of Greek is still in progress: several areas (like information structure and topic-focus articulation) require supplementary work. The annotation of verb valency and co-reference resolution, however, is complete. Our valency lexicon can also be used as a reference for any project that intends to carry on this kind of annotation. 37

The evaluation of the historical information that can be extracted from such a corpus and its uses for research in ancient history is the first and foremost task that awaits us right now. 38

Notes

[1] More details on the composition of the *AGDT* in its latest release and on the formalism used to encode linguistic analysis will be provided in Section 2.1

[2] The word is used in the ancient *Scholia* to Thuc. 1.89.1 [Hornblower 1991, 133]; for a general reading of the passage see most recently Bresson (2010), with many references to the previous literature.

[3] To be more precise, this count refers to treebank *tokens* (see, Section 2 and note below), which do not always correspond to words.

[4] Details of this effort were presented by the staff of Hellespont in a session of the *Digital Classicist — Berlin 2013-4* seminar [Thomas et al. 2013].

[5] For one example, the revolt of Samos and its suppression, see http://gapvis.hellespont.dainst.org/#book/1/event/55

[6] In fact, in this case not even the context is sufficient. This passage is notoriously ambiguous and may support two different interpretations. In the one adopted, Thucydides comments about the Spartans' deficiency in siege warfare (this is the interpretation reflected by the English translations of Jowett 1881 and Crawley 1910). But the Greek text supports also a different reading, based on what we know of the sequel: while the operations dragged on, the Athenians could not bring any decisive aid to the siege and were therefore dismissed; probably, the lack of progresses in the campaign was used as a pretext to hide the political suspicions that the Spartans began to have on their former allies, as Thucydides himself states (see 1.102.3-4). This passage can then be translated accordingly, with the Athenians as the referent of the pronoun *τοῖς* ("the Athenians' deficiency in siege-warfare became evident") and (implied) subjects of *εἷλον* (not even the Athenians could storm the fortress); this interpretation is adopted by Classen (1919), Maddalena (1952) and De Romilly (1958).

[7] Technically, the units of annotation are called *token*; in most cases, a treebank token corresponds either to a word or to a punctuation mark, but there are notable exceptions. A frequent case of mismatch is represented by coordinating conjunctions: the Greek equivalent of the English "neither", for example, is analyzed in the treebank as two distinct tokens, one for the negative particle, one for the coordinating conjunction. Sentence tokenization is based on the punctuation of the edition of Jones and Powell (1942), as digitized in the Perseus Digital Library.

[8] On the PDT see Bohmova et al. (2001). More information on the dependency formalism of AGLDT and IT-Treebank can be found in Passarotti (2009) and Mambrini (2011).

[9] The results of the first experiments with automatic parsing of ancient Greek are discussed by Mambrini and Passarotti (2012).

[10] A more comprehensive introduction, with a lengthy discussion on the annotation of two passages of Thucydides, can be read in Mambrini (2013)

[11] Thus, for example, in the hypothetical English sentence "I want to dance" only "I" and "dance" would be represented as nodes of a tectogrammatical tree. "Dance" would acquire the *volition* deontic modality for wanted/intended events.

[12] A complete list of the functors used in the PDT can be found in the manual for tectogrammatical annotation [Mikulová et al. 2006, Chapter 7].

[13] A single word can have different valencies, which would be recorded in multiple *valency frames*. For example, the English verb "to act" can be used with the meaning "to take action", and in that case it only requires an *actor* as argument; or it can be used with the meaning "to play a role", like in the sentence "the Bank will act as a trustee"; this second frame will include an *actor* (the bank, in the example) and a *predicative complement* (as a trustee). Nouns and adjectives can of course have valency too. Their argument structures (with ellipsis and co-reference resolution) have been annotated in the text, but the entries for nouns and adjectives in the valency lexicon have not been created yet.

Works Cited

Abeillé 2003 Abeillé, Anne, ed. *Building and Using Parsed Corpora*, Dordrecht and Boston: Kluwer Academic Publishers, 2003.

Bamman et al. 2009 Bamman, David, Francesco Mambrini, and Gregory Crane. *An Ownership Model of Annotation: The Ancient Greek Dependency Treebank*. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*. Milan: EDUCatt, 2009, pp. 5-15.

Bohmová et al. 2003 Bohmová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká. *The Prague Dependency Treebank: A Three-Level Annotation Scenario*. In Abeillé 2003, pp. 103-27.

Bresson 2010 Bresson, Alain. *Revisiting the Pentekontaetia*. In *Ombres de Thucydide: La Réception de L'historien Depuis l'Antiquité Jusqu'au Début Du XXe Siècle*, edited by Valérie Fromentin, Sophie Gotteland, and Pascal Payen. Bordeaux: Ausonius, 2010, pp. 383–401.

Cinková et al. 2008 Cinková, Silvie, Eva Hajičová, Jarmila Panevová, and Petr Sgall. *The Tectogrammatics of English: On Some Problematic Issues from the Viewpoint of the Prague Dependency Treebank*. In *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, edited by Joakim Nivre, Mats Dahllöf, and Beáta Megyesi. Acta Universitatis Upsaliensis, 2008, pp. 33–48.

Classen and Steup 1919 Classen, Joannes, and Julius Steup. *Thukydides. Erklärt von J. Classen. Bearbeitet von J. Steup. Vol. 1.*. 5th ed. Berlin: Weidmann, 1919.

Crawley 1910 Crawley, Richard. *Thucydides. The Peloponnesian War. Translated by Richard Crawley*. London and New York: J. M. Dent and Dutton, 1910.

De Romilly 1958 De Romilly, Jacqueline. *Thucydide. La Guerre Du Péloponnèse. Livre I. Texte Établi et Traduit Par J. de Romilly. Vol*. 1. 2nd ed. Paris: Les Belles Lettres, 1958.

Doerr 2003 Doerr, Martin. *The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata*. *AI Magazine*, 24.3 (2003). Available online at http://dl.acm.org /citation.cfm?id=958671.958678.

Haug and Jøhndal 2008 Haug, Dag Trygve Truslew, and Marius Larsen Jøhndal. *Creating a Parallel Treebank of the Old Indo-European Bible Translations*. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*. Marrakech, Morocco: European Language Resources Association (ELRA), 2008, pp. 27–34.

Hornblower 1991 Hornblower, Simon. *A commentary on Thucydides. Vol. I. Book I-III*. Oxford; New York: Clarendon Press, 1991.

Jones and Powell 1942 Jones, H.S., and J.E. Powell. *Thucydides. Historiae*. Clarendon Press, 1942.

Jowett 1881 Jowett, Benjamin. *Thucydides. Vol 1*. Oxford: Clarendon Press, 1881.

LSJ 1940 Liddell, Henry George, and Robert Scott. *A Greek-English Lexicon*. Revised and augmented throughout by. Sir Henry Stuart Jones, with the assistance of Roderick McKenzie. Oxford: Clarendon Press, 1940.

Maddalena 1952 Maddalena, Antonio. *Thucydidis Historiarum Liber Primus. Vol. 2*. Firenze: La Nuova Italia, 1952.

Mambrini 2011 Mambrini, Francesco. *L'Ancient Greek Dependency Treebank. Un nuovo strumento per lo studio della lingua greca*. *Lexis* 29 (2011): 51–70.

Mambrini 2013 Mambrini, Francesco. *Thucydides 1.89-118: A Multi-layer Treebank*. *CHS Research Bulletin*. 1.2 (2013). Available online at http://nrs.harvard.edu/urn-3:hlnc.essay:MambriniF.Thucydides_1.89-118_Multi-layer_Treebank.2013

Mambrini and Passarotti 2012 Mambrini, Francesco, and Marco Passarotti. *Will a Parser Overtake Achilles? First*

Experiments on Parsing the Ancient Greek Dependency Treebank. In *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories (TLT11)*, edited by Iris Hendrickx, Sandra Kübler, and Kiril Simov. Lisbon: Colibri, 2012, pp. 133-44.

Mikulová et al. 2006 *Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank*. Prague: Institute of Formal and Applied Linguistics, 2006. Available online at http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html

Passarotti 2009 Passarotti, Marco. *Theory and Practice of Corpus Annotation in the Index Thomisticus Treebank*. *Lexis* 27 (2009): 5–24.

Romanello and Thomas 2012 Romanello, Matteo, and Agnes Thomas. "The World of Thucydides : From Texts to Artefacts and Back", in *CAA Proceedings 2012*, Amsterdam University Press, 2012, pp. 276-284.

Rydberg-Cox 2002 Rydberg-Cox, Jeffrey. *Mining Data from an Electronic Greek Lexicon*. *Classical Journal* 98.2 (2002): 183-188.

Sgall et al. 1986 Sgall, Petr, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Dodrech: Academia/Reidel Publishing Company, 1986.

Sgall et al. 2004 Sgall, Petr, Jarmila Panevová, and Eva Hajičová.*Deep Syntactic Annotation: Tectogrammatical Representation and Beyond*. In *Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, edited by A. Meyers. Boston: Association for Computational Linguistics, pp. 32-38.

Tesnière 1959 Tesnière, Lucien. *Éléments de syntaxe structurale*. Paris: Klinksieck, 1959.

Thomas et al. 2013 Thomas, Agnes, Alexander Recht, and Karen Schwane. *The Hellespont Project: Integrating Arachne and Perseus in a new Linked Data interface*. Seminar presented at Digital Classicist -- Berlin, 2013-14, abstract, slides and video available at http://de.digitalclassicist.org/berlin/2013/12/17/Thomas-Recht-Schwane, posted on Dec. 17th 2013.