Querying Syntactic Constructions in Ancient Greek Parsed Corpora : A Case Study on the Genitive Absolute in Literature and Documentary Papyri

Vierros, Marja

2022-04-30

Vierros , M & Valentinova Yordanova , P 2022 , ' Querying Syntactic Constructions in Ancient Greek Parsed Corpora : A Case Study on the Genitive Absolute in Literature and Documentary Papyri ' , Classics@ , vol. 20 . <
https://classics-at.chs.harvard.edu/querying-syntactic-constructions-in-ancient-greek-parsed-corpora-a-case-study-or
>

http://hdl.handle.net/10138/351159

unspecified
publishedVersion

Querying Syntactic Constructions in Ancient Greek Parsed Corpora: A Case Study on the Genitive Absolute in Literature and Documentary Papyri

Marja Vierros, Polina Yordanova

1. Introduction

As a distinct and easily identifiable linguistic feature, the genitive absolute construction in Ancient Greek is an excellent example to illustrate how morphosyntactically annotated corpora can be examined for linguistic phenomena and what challenges arise in dealing with heterogeneous data such as papyrological material. [1] By comparing the results obtained from querying two corpora of literary texts (AGDT and Gorman) with those extracted from the PapyGreek corpus of Greek documentary papyri through XSLT-based queries, we explore the different usage and distribution of the construction across time and genre. Additionally, we apply alternative querying processes to the corpora of two major treebanking projects—PROIEL and the automatically parsed Duke-nlp—outlining the pros and cons of the different approaches.

This paper's intention is to investigate one relatively simple and distinct syntactic construction, using different dependency treebanks of Ancient Greek, in order to discover what kinds of questions we are able to answer with the data from the morphosyntactically-annotated corpora that have only recently become available. [2] As the field is in constant movement with the increased production of data, it is important to also evaluate the benefits and limitations of parsed corpora in research. The genitive absolute (GA) construction is an independent,

optional constituent of a clause, and the information its clausal structure provides could be arranged in many other ways as well. It is interesting to study how and in which contexts it is used, in order to say something about, for example, how widespread the GA was in everyday speech and whether it featured predominantly in certain genres or in the work of certain writers—and in what ways they used it. Our own treebank of Greek documentary papyri, which is under development, is examined and used in parallel with larger literary treebanks. We also briefly explore an automatically parsed corpus of papyri, as it provides different benefits and challenges than the above-mentioned semi-manually annotated and vetted corpora.

2. Genitive Absolute—How Is It Used and Is It Always Absolute?

Let us begin by examining how the GA construction is defined by the Encyclopedia of Ancient Greek Language and Linguistics (Buijs 2013):

> A genitive absolute is a construction consisting of, at least, a participle in the genitive case, either sg. or pl., and, usually, a noun in the genitive case agreeing with the ptc. in gender and number. The construction is called 'absolute' because the noun in the genitive case does not perform a syntactic function in the matrix clause; it performs the syntactic function of subject of the participial predicate.

In other words, we are talking about a participial phrase that is independent of the structure of the main clause or the rest of the sentence. [3] It is an optional constituent, and many grammars discuss it within the section of circumstantial participles. An example is from Isocrates 9.56:

> καὶ ταῦτ' ἐπράχθη Κόνωνος … στρατηγοῦντος
>
> And these things were done while Konon was general (literally, 'Konon being the general')

The recently published Cambridge Grammar of Classical Greek defines the GA as, "When the subject of the participle is not a constituent of the matrix clause, it must be expressed separately. In this case, both the participle and its subject are added in the genitive case." It supplements this definition by noting that occasionally the subject is *not* expressed if it can be easily supplied from the

context. [4]

Grammars describe and identify structures used in our sources. Classical Greek sources consist mostly of works of literature, and therefore we primarily study Ancient Greek as an art form and focus less on how it was employed in everyday life. This is where documentary texts, such as inscriptions and papyri, can fill the gap to some extent. Furthermore, lately, the study of linguistic variation used in these sources has experienced a significant gain in popularity. [5] Naturally, neither inscriptions, papyri, nor even graffiti are free from formulaic and educated language use, which needs to be taken into consideration when studying language from written sources only. Documentary papyri consist of different text types, and some reflect everyday speech better than others. For this reason, we wish to explore how a structure like the GA is used in papyri. The underlying assumption is that the—sometimes overflowingly abundant—use of circumstantial participles was rather a feature that made the language seem more educated, literary, or legally valid than a living feature of the language.

Holger Thesleff addressed this question, suggesting that Ancient Greek grammars seem to take the GA as a normal and neutral part of the language. [6] In his study on the use of the GA in Plato's dialogues, he found out that it appears more in the rhetorical than in the colloquial dialogue sections. He concluded that the GA was a stylistic device commonly employed in formal or strict narrative, rhetorical or otherwise formal argumentation, and in various legal and ceremonious contexts—and was, therefore, not an organic part of everyday speech. [7] This is in accordance with Jannaris' statement concerning postclassical Greek that popular speech preferred the simpler and clearer mode of substituting the circumstantial participle with either a prepositional infinitive or, far more commonly, a finite clause, which is either subordinate or co-ordinate to the principal clause. [8]

When studying the postclassical Greek featured in documentary papyri, one should always consult Mayser's grammar. According to Mayser, the GA does not retain all its spectrum in Koine Greek. He also makes the important observation that in papyri, it is used most often in a way that would be unacceptable in Classical Greek and rare in the New Testament (NT); that is, the subject of the GA and the matrix clause coincide. [9] One example from Mayser, which is also in our corpus, is P. Cair. Zen. 2 59245, 1:

ἀπελθόντος μου ἀπὸ σοῦ κατέλαβον τοὺς γεωργο[ὺς]…
ἀνακεχωρηκότας…

After I left you, I discovered that the peasants … had fled…

Here, the genitive absolute construction has the first-person singular pronoun (μου) as the agent—the subject of the absolute construction. However, the matrix clause predicate is also in the first-person singular (κατέλαβον) and refers to the same person. In Classical Greek, one would instead expect a circumstantial participle in the masculine singular nominative (ἀπελθών) to agree with the subject of the matrix clause predicate (*participium coniunctum*).

The "non-absolute" usage has been noted in Hellenistic Koine by other scholars as well. [10] The definition of non-absolute GA covers all instances where the subject of the GA *appears* also in the matrix clause. Yet it is extremely rare that the subjects of the GA and the matrix clause coincide in Septuagint or NT. [11] Fuller argues that the cases of GA in which the subject of the participle also plays a syntactic role in the matrix clause appear so widely in Hellenistic Greek that they should not be called "irregular GAs" or described as "violating" the rules of classical Greek GA, as it is done in most grammars. [12] In her view, the use of GA in NT and Hellenistic Greek is not about being absolute but is used to draw the reader's attention to certain background information with more prominence than other circumstantial participles do—i.e. GA acts as an essential frame to the information in the main clause. Often, it indicates the change of scene or location, especially when related to time, as the genitive of time is historically behind the whole construction. [13] She also suggests that the GA should be called differently: e.g., a Genitive Construction (GC). [14] In this article, we use the name genitive absolute, since "genitive construction" could be used for many other constructions involving the genitive as well. It is, therefore, justified to use a traditional name easily understood by anyone who has ever read Ancient Greek grammar.

3. Parsed Corpora and Querying

There exist a number of parsed corpora for texts in Ancient Greek, many of which follow roughly the same formalism of dependency grammar, originating from the Prague Dependency Treebank. [15] They have been presented and discussed previously, [16] so only a short introduction is needed here.

In its most basic form, a parsed corpus consists of texts that have been divided into sentences based on punctuation. A sentence forms one syntactic tree, where each token—word or punctuation mark—is marked with its linear place in the sentence (word-id), its form and lemma, and its syntactic role in the sentence as well as its head (governor). In addition, in all treebanks discussed here, the morphological analysis of each word is added. In the syntactic tree, each word, except one, has one head (parent). The main predicate—or occasionally a coordinator—acts as the root of the sentence (head="0"). A token can have one or more dependents (children), but it does not have to govern any words—see, for example, the standard treebank structure in XML in the next section. The morphological parsing and the syntactic roles for Ancient Greek are described in guidelines—see more below—which each annotator tries to follow to the best of their ability. However, since language is not a simple mathematical calculation, some differences can be generated due to different interpretations of some words or their roles in the sentence.

Below are presented three corpora for Ancient Greek that we have queried in the Kiln XSLT platform (discussed below) for this article. Two other corpora —namely PROIEL and Duke-nlp—are not included in our Kiln queries; they have their own querying interfaces, and they will be discussed at the end of the article. Only Duke-nlp has been automatically parsed; all others have had human annotators and also been subject to, in most cases, a review process.

3.1 AGDT

The earliest treebank for Ancient Greek and Latin is the Ancient Greek and Latin Dependency Treebank, [17] and the acronym AGDT is used for the Greek part. We are using the latest release (2.1), excluding Homer. [18] As such, it contains 321,829 tokens from thirteen authors. There are tragedies—by Aeschylus and Sophocles—, epics—by Hesiod and Pseudo-Homer—, and prose—by Aesop, Apollodorus, Athenaeus, Diodorus Siculus, Herodotus, Plato, Plutarch, Polybius, and Thucydides. [19] The annotations have been made semi-manually by several people with the help of the Arethusa tool provided by the Perseids project. [20] The annotations follow the Guidelines for Ancient Greek Dependency Treebank 2.0. [21] A few of them also include more detailed annotation than described above—the so-called advanced syntactic layer/semantic layer. [22]

3.2 Gorman Trees

A large treebank corpus is annotated by one person, Vanessa Gorman. [23] Its c. 600,000 tokens represent a great variety of prose, mainly by historical and rhetorical writers. Gorman has contributed also to the AGDT corpus, and therefore some parts of these two corpora are duplicates—we have not excluded any parts because of this but rather consider the Gorman corpus as one entity and AGDT as another. Guidelines used by Gorman are 1.1, [24] where some syntactic labels that no longer exist in 2.0 are still used, but the differences between the two are insignificant to our genitive absolute queries.

3.3 PapyGreek

The corpus of the PapyGreek project [25] is a continuation of the Sematia-corpus. [26] The data from old Sematia is currently being migrated to the new platform PapyGreek while also being re-checked to follow our guidelines. [27] There are minor differences in tokenization as well as text division compared to the old Sematia. [28] The basic idea of the Sematia corpus of annotating two versions (or layers) of each document is kept intact: we produce the original corpus, where only the preserved forms are annotated, and the regularized corpus, in which the texts are fully annotated according to the editorial supplements and regularizations. In other words, the regularized corpus represents idealized versions of the texts, whereas the original corpus represents the reality with all its linguistic variation and fragmentariness. Both corpus layers still contain many more gaps within the sentences than the literary corpora discussed above. This also means that sometimes a syntactic tree is not so clear in its parent-child ratios: we may have several branches of the tree fallen into the ground, so to say, if the syntactic links are missing due to gaps in the papyrus. For this reason, the search results for the original versus regularized corpus can yield different figures; some syntactic constructions may not appear in the original at all or they may appear to have fewer components.

The types of texts in the corpus have not yet been chosen systematically. That is to say, the representativity of the corpus will develop only in the later stages; we have concentrated on annotating certain archives. There are texts from the Zenon archive, [29] the Memphis Katochoi archive, [30] the archive of Athenodoros, [31] and letters written on potsherds found in the military garrison

Mons Claudianus (2nd cent. CE). Version 1.01 is focused on the Hellenistic period (ca. 75% come from the period BCE) and, with text types, towards letters and petitions.

3.4 Queries in Kiln Platform

The data used for the study of GA across the three corpora is obtained through queries performed in the Kiln platform, [32] which was customized for handling morphosyntactically annotated corpora.

Kiln is a multi-component XSLT-based platform for the manipulation and publication of XML documents, developed at King's College London. Thanks to its high customizability, it has been used as the base in the online publication of over 50 projects. In its incarnation as a tool for querying treebanked corpora, it was developed as part of Polina Yordanova's doctoral research with extensive support from Jamie Norrish, one of the platform's lead architects.

In treebanking, the morphosyntactic annotation performed on the texts is recorded in an XML document, where the text is divided into 'sentence' elements and the words within each sentence are represented by individual 'word' elements in word order. Each 'word' element has attributes containing the information regarding its morphology, its syntactic function, and its relation to other words on the tree (see Figure 1 for the display of a syntactic tree and Figure 2 for the XML document). In the Kiln platform, this XML content can be enriched with additional annotations that mark a particular feature of a given linguistic phenomenon, which has been checked and corrected for consistency, and restructured or otherwise manipulated in order to facilitate querying, while the input documents remain unchanged. A built-in templating system and pipelines allow for the visualization of the results in a browser window.

When querying treebanks for any feature, a major challenge is traversing the tree's structure to find dependencies between the words. In the standard treebank XML structure, these are recorded through the @id and @head attributes, representing respectively the word's place in the sentence in word order and its immediate governor. Often, however, we are not interested in the head—i.e., in the immediate parent element—of a given word (in cases of coordination, for example) but in the ancestors. Therefore, as a first step in preprocessing the document for querying, the XML is restructured to more

closely represent the actual tree hierarchy by making the head 'word' parent elements of their dependents. This disrupts the linear structure of the XML and takes advantage of the possibility to query directly using the ancestor-dependent axis, without the need to check each time the words @id and @head to establish the dependency relations between them (see Figure 3).

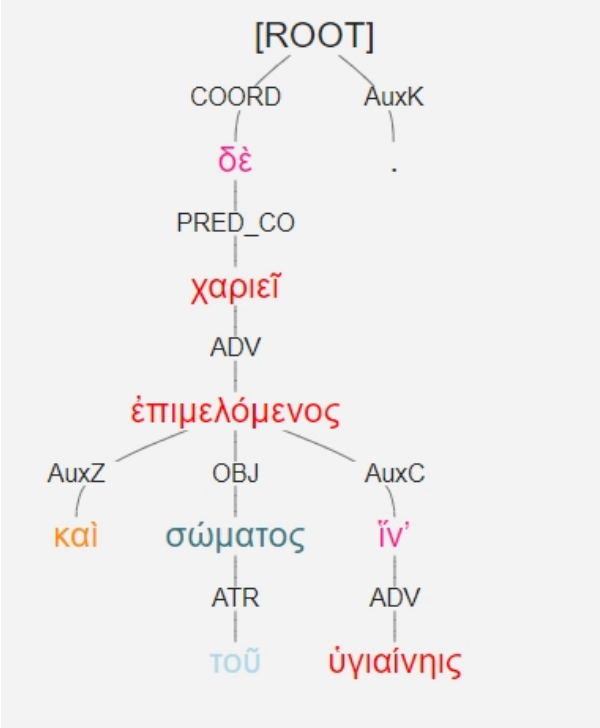


Figure 1. Tree structure of the annotated sentence.

```
▼<sentence document_id="https://sematia.hum.helsinki.fi/edit/644" id="6" span="" subdoc="">
   <word form="χαριεῖ" head="2" id="1" lemma="χαρίζω" postag="v2sfim---" relation="PRED_CO"/>
   <word form="δὲ" head="0" id="2" lemma="δέ" postag="c--------" relation="COORD"/>
   <word form="καὶ" head="6" id="3" lemma="καί" postag="c--------" relation="AuxZ"/>
   <word form="τοῦ" head="5" id="4" lemma="ὁ" postag="l-s---ng-" relation="ATR"/>
   <word form="σώματος" head="6" id="5" lemma="σῶμα" postag="n-s---ng-" relation="OBJ"/>
   <word form="ἐπιμελόμενος" head="1" id="6" lemma="ἐπιμελέομαι" postag="v-sppemn-" relation="ADV"/>
   <word form="ἵν'" head="6" id="7" lemma="ἵνα" postag="c--------" relation="AuxC"/>
   <word form="ὑγιαίνῃς" head="7" id="8" lemma="ὑγιαίνω" postag="v2spsa---" relation="ADV"/>
   <word form="." head="0" id="9" lemma="punc1" postag="u--------" relation="AuxK"/>
 </sentence>
```

Figure 2. Standard linear treebank XML structure.

```
▼<sentence document_id="https://sematia.hum.helsinki.fi/edit/644" id="6" span="" subdoc="">
  ▼<word form="δὲ" head="0" id="2" lemma="δέ" postag="c--------" relation="COORD">
    ▼<word form="χαριεῖ" head="2" id="1" lemma="χαρίζω" postag="v2sfim---" relation="PRED_CO">
      ▼<word form="ἐπιμελόμενος" head="1" id="6" lemma="ἐπιμελέομαι" postag="v-sppemn-" relation="ADV">
         <word form="καὶ" head="6" id="3" lemma="καί" postag="c--------" relation="AuxZ"/>
        ▼<word form="σώματος" head="6" id="5" lemma="σῶμα" postag="n-s---ng-" relation="OBJ">
           <word form="τοῦ" head="5" id="4" lemma="ὁ" postag="l-s---ng-" relation="ATR"/>
         </word>
        ▼<word form="ἵν'" head="6" id="7" lemma="ἵνα" postag="c--------" relation="AuxC">
           <word form="ὑγιαίνῃς" head="7" id="8" lemma="ὑγιαίνω" postag="v2spsa---" relation="ADV"/>
         </word>
       </word>
     </word>
   </word>
   <word form="." head="0" id="9" lemma="punc1" postag="u--------" relation="AuxK"/>
 </sentence>
```

Figure 3. Hierarchical tree structure after preprocessing.

Having standardized and restructured the XML documents, we created additional annotations in order to make the phenomenon more easily discoverable. In devising the query process for genitive absolute constructions, we, first and foremost, had to establish rules regarding the possible configurations of dependencies between the participants in a GA construction within the treebanking framework. As a circumstantial participle, the head of the GA is marked with the syntactic label ADV, and its morphological analysis contains the information that it is a verb, participle, and in the genitive case. Its required dependent is the subject of the participle and is marked with syntactic label SBJ; the only morphological requirement needed is the genitive case. [33] Since the participle and the genitive agent can participate in 'direct parent'–'direct child,' 'indirect ancestor'–'indirect dependent' relations, or, if a coordinator is involved, even be 'sibling' elements, "valid paths" were defined from each component of the construction to each other participant, with coordinators factored in as possible "bridges" between them (see below examples of the four different types). Each word satisfying the morphological requirements and able to reach another potential participant in the construction through a valid path of dependency is given a @group attribute, the value of which is taken to be the value of the @id of the first genitive agent in word order (see Figure 4).

Figure 4. Tree fragment "τοῦ πατρὸς ζῶντος καὶ τῆς μητρός" with annotated genitive absolute participants.

```
<constructions>
  <construction sentence="1" type="2" group="3" masculine="1" feminine="1" noun="2" number-of-agents="2" number="s"
    constituents-order="mixed" order-in-sentence="vc"/>
</constructions>
```

Figure 5. A summary element for the construction above.

A rigorous and detailed test suite has been implemented to ensure that the additionally-annotated preprocessed files are neither falsely identifying genitive absolute constructions, nor omitting any actual constructions from the count. This workflow allows us to have a high certainty level for the obtained results.

4. Parsed Corpora and Genitive Absolute

4.1 Overview

In this section, we will examine the use of the GA construction in two—partly overlapping—corpora of treebanked Ancient Greek literature and then compare the results with our treebanked corpus of Greek documentary papyri. First, some general counts from these corpora in Table 1.

Counts	*AGDT*	*Gorman*	*PapyGreek: orig (reg)*
1. Tokens (all)	321829	605779	44309 (44098)
2. Tokens (minus punctuation, gaps, artificial tokens)	281675	540438	37215 (37283)
3. Sentences	18417	25731	3102 (3103)
4. Sentences with GA	1200	2773	131 (142)
5. Total number of GA	1427	3338	189 (210)
6. % of sentences with one or more GA (of all sentences)	6.52	10.78	4.22 (4.58)
7. % GA / Number of sentences	7.75	12.97	6.09 (6.77)
8. % GA / Number of all tokens	0.44	0.55	0.43 (0.48)

Table 1. The sentence and token counts for the corpora and counts for the genitive absolute construction in the corpora as a whole.

4.2 Genitive absolute in the literary corpora

When we study the appearance of the genitive absolute construction in the AGDT and Gorman corpora author by author, it is clear that epic and tragedy have the least number of occurrences (see Table 2). This indicates that the genitive absolute was seldom considered to be a suitable linguistic construction in elevated, orally delivered poetry. It would be a stretch to say that it speaks directly to its use in spoken language.

Author, (AGDT corpus)	*1. Tokens*	*3. Sentences*	*4. Sentences with GA*	*5. Total number of GA*	*6. % of sentences with one or more GA*	*7. % GA / Number of sentences*
Aeschylus	48449	3958	45	46	1.14%	1.16%

Hesiod	19284	1183	20	22	1.69%	1.86%
Ps-Homer	3968	255	5	5	1.96%	1.96%
Sophocles	50094	4001	53	55	1.32%	1.37%

Table 2. Epic and tragedy writers, and their use of the genitive absolute in archaic and classical Greek.

The second group to be taken as its own entity is formed by the orators or rhetorical writers of the classical period. Some stylistic variations can be observed: for example, Demosthenes favors the GA much more than other authors (see Table 3). As rhetorical texts were also meant to be orally delivered, their use of the GA indicates that it was considered a possible, although not common, spoken feature in the courtroom.

Author, date (Gorman corpus)	*1. Tokens*	*3. Sentences*	*4. Sentences with GA*	*5. Total number of GA*	*6. % of sentences with one or more GA*	*7. % GA / Number of sentences*
Antiphon, 5th BCE	16433	764	49	52	6.41%	6.81%
Lysias, 5th/4th BCE	22122	971	63	73	6.49%	7.52%
Demosthenes, 4th BCE	58038	2134	223	281	10.45%	13.17%
Aeschines, 4th BCE	15971	678	36	41	5.31%	6.05%

Table 3. Orators and their use of the genitive absolute.

The largest group of texts comes from the prose writers (see Table 4). The historians clearly use the GA more than others, although the early ones—Herodotus and especially Xenophon—are behind Demosthenes in terms of orators. [35] Thucydides and Aesop are both around the same level as Demosthenes. Plato's dialogues have quite low numbers, which reflect one of the starting points of this article. In this kind of corpus query, however, we cannot automatically reach a conclusion similar to Thesleff's notion that the GA appears only in the more rhetorical parts since those parts are not indicated as such in the corpora—naturally, we can go and examine the GA constructions in the text for further qualitative analysis. Historians of the Hellenistic and Roman periods are among the top users of GA in the corpora, while the highest numbers are found in Diodorus Siculus. Due to the selection of writers in the corpora, we cannot easily say if the percentages tell us about the writing style of the genre only or if there also is a chronological aspect to be considered. However, the lower percentages in Pseudo-Apollodorus (*Bibliotheca*) that contain mythological descriptions and in Athenaeus (*Deipnosophistai*), whose

work is a mixture of genres that includes citations from other authors, suggest that it is the historical genre that most favors the GA in its stylistic repertoire. A desideratum would be more annotated corpora from representatives of other genres from the Hellenistic and Roman periods (see below on the PROIEL corpus of New Testament and some late Greek).

We will now turn to the documentary papyri to provide a picture of the Hellenistic and Roman periods.

Author, date (Gorman corpus)	*1. Tokens*	*3. Sentences*	*4. Sentences with GA*	*5. Total number of GA*	*6. % of sentences with one or more GA*	*7. % GA / Number of sentences*
Aesop (AGDT)	5221	366	44	47	12.02%	12.84%
Ps.-Xenophon, 5th BCE (Gorman) [36]	3723	170	4	4	2.35%	2.35%
Herodotus, 5th BCE (Gorman/AGDT) [37]	33150/33102	1555/1555	132/134	154/156	8.49%/8.62%	9.9%/10.03%
Thucydides, 5th BCE (Gorman/AGDT)	32344/25266	1204/942	124/101	151/127	10.3%/10.72%	12.54%/13.48%
Xenophon, 4th BCE (Gorman)	57903	2811	175	205	6.23%	7.29%
Plato, 4th BCE Apology (Gorman)	10457	481	17	20	3.53%	4.16%
Plato, 4th BCE Eythyphro (AGDT)	6349	426	5	5	1.17%	1.17%
Aristotle, 4th BCE (Gorman)	19867	871	35	42	4.02%	4.82%
Polybius, 2nd BCE (Gorman/AGDT)	105693/28271	3816/1001	648/187	803/232	16.98%/18.68%	21.04%/23.18%
Diodorus Siculus, 1st BCE (Gorman/AGDT)	25692/25660	991/991	245/244	308/307	24.72%/24.62%	31.08%/30.98%
Dionysius Halicarnassus, 1st BCE (Gorman)	30312	1067	135	162	12.65%	15.18%
Josephus, 1st CE (Gorman)	24987	1039	113	131	10.88%	12.61%
Plutarch, 1st/2nd CE (Gorman/AGDT)	37203/22124	1479/865	230/163	287/203	15.55%/18.84%	19.41%/23.47
Apollodorus-Ps., 1st/2nd CE (AGDT)	1265	51	3	3	5.88%	5.88%
Appian, 2nd CE (Gorman)	25665	966	204	248	21.12%	25.67%

Athenaeus, 2nd/3rd CE (Gorman/AGDT)	86219/45585	4734/2525	340/175	376/195	7.18%/6.93%	7.94%/7.72%

Table 4. Prose writers (in roughly chronological order) and their use of GA

4.3. Genitive absolute in PapyGreek corpus

As seen in Table 1, the percentage of sentences containing the GA construction from all sentences in the corpus is 4.2% for the original layer, and somewhat higher in the regularized layer, as we have more fragmentary or missing branches in the original. The higher percentage in the regularized layer may indicate that the construction is often present in a formulaic phrase that has been easy for the editors to supplement. Nonetheless, with such a small corpus, one should never take these percentages as very precise indicators. However, when we split the corpus down to different text types, one striking tendency can be seen: the corpus consists of mostly letters and petitions, and it is very clear that the petitions contain the majority—roughly 65%—of the GA constructions (see Table 5). [38] Moreover, 80% of petitions contain at least one GA construction. [39] From 278 letters, on the other hand, only 36 contain one or more GA construction. [40] It is, therefore, quite safe to say that the language in petitions —legal and formulaic but also narrative—favors the GA construction, but private letters usually avoid it. One can expect that when the number of contracts rises in the corpus, they have somewhat higher numbers of the GA than the letters—at the moment we see that in texts not yet in the released corpus. [41]

Text type	*Text count*	*Texts with GA (% of texts within type)*	*GAs within type (% of GAs within corpus)*
Letter	278	36 (13%)	53 (28%)
Petition (with attachments)	59	48 (81%)	123 (65%)
Contract	12	1 (8%)	1 (0.5%)
Other types	13	6 (46%)	9 (4.7%)
Text type not defined	4	1 (25%)	3 (1.6%)

Table 5. Text type and GA (PapyGreek orig)

4.4. Comparisons between the literary and PapyGreek

corpora

Many different details about GA constructions can be deciphered with the annotated corpora. We will go through some of them and compare if we find further differences in the papyrological language when compared to the literary corpora.

4.4.1 Agent

The agent—i.e., the subject of the participle—is expressed in different ways. It is easy to extract the part-of-speech (POS) of the agent: noun (proper or common), pronoun, adjective, or verb (i.e., another participle), etc. In the PapyGreek orig corpus, there are 198 agents in the 189 GA constructions and 48% are pronouns. Nouns constitute 43% (24% proper nouns), and adjectives and verbs the rest. When we consider how the adjective or the verb works as an agent, it is clear that they are substantivized—e.g. τῶν δὲ τῆς μητρὸς φίλων ἀναπεισάντων ἡμᾶς… 'when the friends of our mother had persuaded us…' upz.1.19 and …τοῦ δὲ πωλοῦντος μὴ βουλομένου ἀποδόσθαι ἐξ ὧν ἔθος πᾶσι πωλεῖν, ἀλλὰ βουλομένου ἄλλα εὐτελέστερά μοι δοῦναι… 'but the seller refused to sell the ones he normally sells to everyone, wanting to give me others of lower quality' upz.1.12). Thus, we could, in fact, count all of those as nouns as well. It is also uncertain how often an annotator has changed the POS from an adjective, which the system offers—e.g., φίλος—for a noun when it is used as a noun; this is not encouraged in the guidelines.

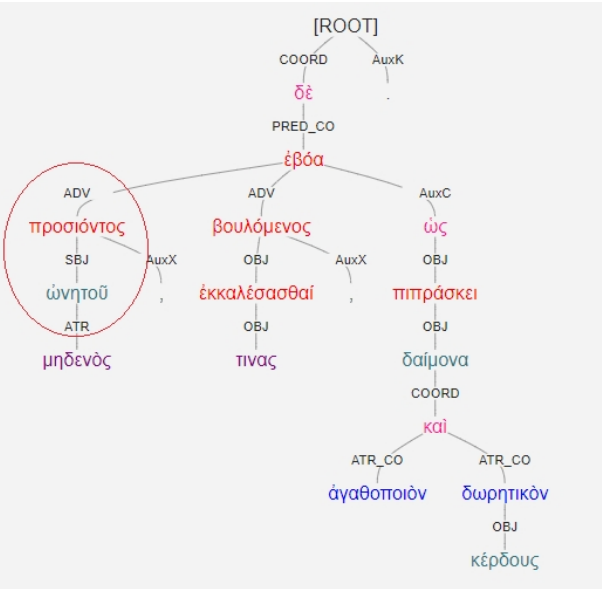
It is worth noting that the agent POS in the AGDT and Gorman corpora constitute only between 20 and 22% of pronouns, respectively. In both, the nouns form ca. 70%. There is, naturally, variation between different authors: some use noun agents more than 70%—e.g., Hesiod (90%), Aeschylus (83%), Diodorus Siculus (79%), Thucydides (73%), Plutarch (73%), and Herodotus (75%)—whereas others use them under the 70% and come closer to the amount in papyri—such as Aesop (53%), Lysias (56%) Polybius (65%), Sophocles (43%), and Xenophon (60%). Although Plato seems to significantly favor pronoun agents at about 60–80%, it is difficult to make decisive conclusions as there are only two dialogues composed by him.

It is also possible to extract the linguistic gender—masculine, neuter, or feminine—from the agents. This could help us in studying the absoluteness of the GA (for which see also below), but as such, it does not speak volumes about

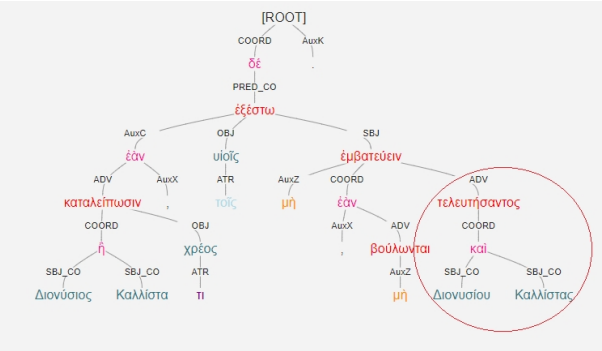
the agents' biological gender, as many common nouns are used as agents as well—e.g., τοσούτου χρόνου ἐπιγεγονότος 'when so much time had passed' upz.1.59. Pronouns, for that matter, can also be neutral—τούτου δὲ γενομένου 'after this happened' upz.1.17. The counts should distinguish masculine and feminine pronouns from proper names and words like 'mother', 'father', etc., in order to get an idea of the sexual gender of the agents in GA constructions.

4.4.2 Single vs co-ordinated governors and agents

For the purpose of devising queries that encompass all instances of the GA, we created a typology based on the possible dependency relations between the participles and their subjects (agents). Four types of GA were defined (see Figure 6): type 1) single participle governing a single agent; type 2) single participle governing co-ordinated agents; type 3) co-ordinated participles sharing a single agent; type 4) co-ordinated participles governing shared co-ordinated agents.

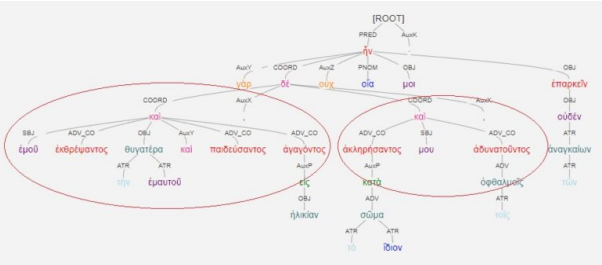


> μηδενὸς δὲ ὠνητοῦ προσιόντος, ἐκκαλέσασθαί τινας βουλόμενος, ἐβόα ὡς ἀγαθοποιὸν δαίμονα καὶ κέρδους δωρητικὸν πιπράσκει. [42]
>
> Since no buyer approached, wanting to call someone forth, he cried that the daemon is beneficent and bestowing profits.



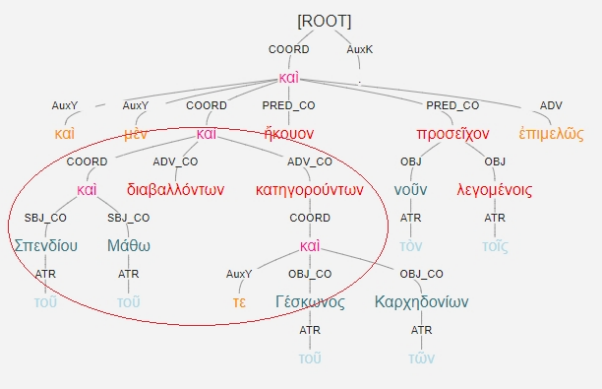
ἐὰν δέ καταλείπωσιν Διονύσιος ἢ Καλλίστα χρέος τι, ἐξέστω τοῖς υἱοῖς μὴ ἐμβατεύειν, ἐὰν μὴ βούλωνται τελευτήσαντος Διονυσίου καὶ Καλλίστας. [43]

And if Dionysios and Kallista would leave behind any debt, may it be possible for the sons not to come into possession, if they do not wish so, after Dionysios and Kallista pass away.



ἐμοῦ γὰρ ἐκθρέψαντος τὴν ἑμαυτοῦ θυγατέρα καὶ παιδεύσαντος καὶ εἰς ἡλικίαν ἀγαγόντος, ἀκληρήσαντος δέ μου κατὰ τὸ ἴδιον σῶμα καὶ τοῖς ὀφθαλμοῖς ἀδυνατοῦντος, οὐχ οἷα μοι ἦν ἐπαρκεῖν τῶν ἀναγκαίων οὐδέν· [44]

For though I raised her, my own daughter, and educated her and brought her to maturity, when I was stricken with bodily ill-health and was losing my eyesight, she was not disposed to furnish me with any of the necessities of life. [45]



καὶ τοῦ μὲν Σπενδίου καὶ τοῦ Μάθω διαβαλλόντων καὶ κατηγορούντων τοῦ τε Γέσκωνος καὶ τῶν Καρχηδονίων ἤκουον καὶ προσεῖχον ἐπιμελῶς τὸν νοῦν τοῖς λεγομένοις. [46]

When Spendius and Mathos began to traduce and accuse Gesco and the Carthaginians, they were all ears and listened with great attention. [47]

Figure 6. Four types of GA.

4.4.3 Word Order

With treebanked data, we can also study in what order the words appear. For the GA construction, we have looked at two features: first, the order of the constituents within the construction—that is, if the agent is placed before or after the participle; and second, whether the GA appears before or after the predicate of the main clause (the matrix verb). For the first parameter, the papyrological data is very similar to the AGDT and Gorman corpora. In all, there is a tendency to place the agent before the participle—a little bit more in papyri than in literature: ca. 62% vs. 57% 'agent-participle' and ca. 34% vs. 39% 'participle-agent'—and some instances of coordinated agents or participles where the order is mixed (ca. 4% for both types of corpora). In the second count, the GA construction is placed before the matrix verb both in the papyrological and literary corpora (ca. 60%). This feature is good to include in future studies of the absoluteness of the GA since the non-absolute constructions are said to be more common when the GA precedes the matrix clause.

5. Other Corpora and the Genitive Absolute

5.1. PROIEL

The parallel corpus of the New Testament for several ancient Indo-European languages includes also the Greek New Testament. [48] In addition, the PROIEL project has included parts of Herodotus' *Histories* with a larger sample than the AGDT/Gorman corpus and a postclassical work, Sphrantzes' *Chronicles* from the 15th century (Table 6). The annotation framework is similar but not identical to the AGDT formalism. [49] The data is available in different formats —e.g., xml and conll) for scholarly use—but it can also be queried directly in the Norwegian CLARIN's *Infrastructure for the Exploration of Syntax and Semantics* (INESS). [50]

PROIEL corpus	*tokens*	*sentences*	*GA total*	*GA % (of sentences)*
Hdt	85080	5446	343	6.30%
NT	140763	11261	240	2.13%
Chron	24612	976	140	14.34%

Table 6. PROIEL corpora and the use of GA.

Table 6 gives the number of GA constructions in these corpora. [51] Interestingly, Herodotus' percentage is somewhat lower than what we saw in the AGDT/Gorman corpora above. The larger PROIEL corpus includes not only the first book of *Histories*, which is also in AGDT/Gorman, but also part of book four and books five through seven, thus showing that the beginning of *Histories* appears to favor slightly the use of the GA. The New Testament writers use the GA very sparingly, almost as little as the epic and tragedy authors. Surprisingly, the late author Sphrantzes uses the GA the most; as a historian, he is obviously following the footsteps of the Hellenistic historians, using the GA approximately with the same frequency as Polybius, Dionysius Halicarnassus, and Plutarch. He also used types 2 and 3 of the GA proportionally more than Herodotus or NT writers, although type 1 was the most popular of all.

5.2. DUKE-NLP

Alek Keersmaekers and the Pedalion project in Leuven are conducting an important and interesting project concerning automatic morphological and syntactic parsing of Ancient Greek. They have a search interface called DendroSearch, which at the moment includes, as a preset, many different manually annotated corpora—e.g. Pedalion Treebanks, PROIEL, AGDT, and

Gorman trees, as well as older Sematia corpus. [52] From our point of view, a very exciting part is the automatically parsed corpus of documentary papyri —Duke-nlp—consisting of most papyri available in the papyri.info. [53] The files can be downloaded and queried with DendroSearch—the Duke-nlp data is divided into several smaller subsections by text types. This provides a large quantity of data (ca. 4.6 million tokens) for linguistic research on papyrological material. However, given that the data is automatically parsed, we must allow the results to include more noise than the manually annotated corpora.

We decided not to incorporate the Duke-nlp data into the Kiln platform, but rather experiment querying the GA in the Duke-nlp files with DendroSearch because it seemed useful to test the query engine provided by the creators of the data and since some inconsistencies (see below) would have made the data incomparable with the ones we tested with Kiln. [54] The results brought us a large number of possible GA constructions. Since the searches were not done identically with the Kiln queries and since we do not have all the same data for them (such as the sentence counts), and sometimes the automated parsing caused false-positive results, [55] Table 7 should be read with due caution. However, we can say something on the basis of those results: the overall percentage of GA (NB. of *tokens*, not sentences) is, at 0.28%, lower than what we have in the three corpora seen in point 8 of Table 1, despite the false positives. It is more useful, however, to compare the figures within the corpus itself. The highest percentage—nearly 1%—comes from the file named "administration"; the one called "declarations" is in second; "contracts," "pronouncements," and "reports" score more than the letters. This is in line with our earlier observation on the lesser use of the GA in letters. A noteworthy but also expected feature came out when combining a lemma constraint to the search; as mentioned above (note 44), the regnal year dating is expressed with a GA construction. In the query of instances in which the participle is of the lemma βασιλεύω, almost half of the type 2 GA numbers were regnal dating formulae in contracts, but not so much in other document types, and the number was especially low in letters. In the type 1 GA, the lemma had a significantly smaller effect in all document types.

Duke-nlp corpus (several files joined)	*tokens*	*GA type 1 (+3?)*	*GA type 2* [56]	*GA% (of tokens)*
Contracts (1+2+3)	1030574	3156	487	0.353%
Pronouncements	86993	259	26.5	0.328%
Declarations (1+2)	525445	2618	162.5	0.529%

Reports	284799	914	61	0.342%
Administration	16902	164	3	0.988%
Accounts	37526	20	1	0.056%
Labels	20755	3	0	0.014%
Receipts (1+2)	721445	751	66.5	0.113%
Letters (1+2+3)	841466	1840	93	0.230%
Lists (1+2+3+4)	1310900	742	22.5	0.058%
Paraliterary	14543	11	0	0.076%
Other	235800	527	57.5	0.248%

Table 7. Duke-nlp corpora and the use of GA

To conclude, the Duke-nlp is a great source due to its large scale. With the help of this corpus, it was possible to find a multitude of GA constructions from the documentary papyri. Unfortunately, the automated parsing is not (yet) perfect. The false positives resulting from incorrect automated parsing require the checking of results to prune those out or even to get an idea of how common or rare they are. [57] The other downside is, of course, that we are unaware of how many GA's are missing due to possible false parsing. The aspiration is to have more manually corrected data to improve the automatic parser—and this is where we hope that PapyGreek corpus can also help in the future.

6. Conclusions and Discussion

The existing morphosyntactically annotated corpora of Ancient Greek tell us that the GA construction was used in different frequencies in different text types. It was less used in tragedy and epic poetry than in prose, and in prose the rhetorical and especially historical writers used it—however, more treebanks from different writers would enrich the image. This suggests that it was one element in the stylistic repertoire of historical narrative writing. The documentary papyri corroborate this image since the construction is most common in petitions, which mainly consist of a narrative of some event where the petitioner has been maltreated. It is also worth noting that letters do not use this construction as often; the corpus naturally includes letters of a different nature—some are more private and mundane, and others more official in nature. The short, everyday letters from the military post at Mons Claudianus contain only one GA (in the treebanked corpus). Therefore, it seems warranted to say that everyday letter writing had no place for the GA—and possibly the same was

the case in everyday speech; more elaborate, official, or administrational letters could contain a GA every now and then.

The function of the GA is frequently to set the time when some event happened, as mentioned e.g., by Fuller (2006), and that is evident also in the frequent use of the GA in datings and other time-related settings. Quite often, we find GA constructions marking a transition from one event to another—of the type τούτου δὲ γενομένου 'after this happened'. Since our study was a preliminary one, we did not delve deeper into the contexts in which the GA was used, but some additional study could well enter into classifying the lemmata of the participles and the matrix verb in order to do that.

An important result of this preliminary study is that it is relatively easy to query a syntactic construction across the parsed corpora we have at our disposal at the moment, even if some differences exist in their annotation styles, output files, or search engines. [58] Naturally, evaluating the results is always important before making far-reaching conclusions. A multitude of aspects could still be studied about the GA using treebanked data. In addition to the semantic context mentioned above, one could examine: e.g. the tense of the participle and the matrix verb and their relation to each other; or the distance between the agent and the participle, and the distance between the GA and the matrix verb; or whether certain combinations of part-of-speech in the GA elements are especially commonly used by certain authors, which could, for example, act as a stylometric indicator for an author. And, certainly, there are many other aspects we cannot even imagine at this moment.

There were some limitations in this treebank-based study too. The question about the absoluteness of the GA—i.e., does the agent of the GA perform some role in the matrix clause— would be difficult to query. Even the extreme case of the subjects coinciding, which Mayser stated to be very common in papyri, could only be inadequately studied in the present treebanks. We could generate a search where the person, number, and gender of the matrix verb and the agent are compared, but that would only give us part of the answer: those cases where the subjects clearly differ from one another. Then, we would still be left with a group of instances where the subjects are in the same person, number, and gender but only reading and understanding the sentence in its context would tell us if the subjects were identical in both cases.

There is still a strong demand for more manually annotated treebanks for further

enhancing the automatic syntactic parsing, especially in the case of documentary papyri. The Duke-nlp corpus is a huge advancement compared to the situation we had only some years ago when no proper computational syntactic queries could be performed for the papyrological data at all. The PapyGreek project aims to provide more vetted data for the endeavor of perfecting automated syntactic parsing. In general, it can be said that the future looks bright in this respect, as more annotations are produced with different methods, and also more exact metadata on text types and writers are added to the corpora.

Bibliography

Bamman, David, and Gregory Crane. 2011. "The Ancient Greek and Latin Dependency Treebank." In *Language Technology for Cultural Heritage*, ed. C. Sporleder, A. van Den Bosch, K. Zervanou, 79–89. Berlin and Heidelberg.

Buijs, Michel. 2013. "Genitive Absolute." In *Encyclopedia of Ancient Greek Language and Linguistics* (EAGLL), ed. G. K. Giannakis. Leiden.

Celano, Giuseppe G.A. 2019. "The Dependency Treebanks for Ancient Greek and Latin." In *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, ed. Monica Berti, 279–297. De Gruyter Saur. DOI: https://doi.org/10.1515/9783110599572.

———. 2014a. "A computational study on preverbal and postverbal accusative object nouns and pronouns in Ancient Greek." *Prague Bulletin of Mathematical Linguistics* 101:97–110.

———. 2014b. *Guidelines for the Annotation of the Ancient Greek Dependency Treebank 2.0.* https://github.com/PerseusDL/treebank_data/edit/master/AGDT2/guidelines (last access 2019.01.31).

Emde Boas, Evert van, Albert Rijksbaron, Luuk Huitink, and Mathieu de Bakker. 2019. *The Cambridge Grammar of Classical Greek*. Cambridge.

Evans, Trevor V. and Dirk D. Obbink, eds. 2010. *The Language of the Papyri*. Oxford.

Fuller, Lois K. 2006. “The ‘Genitive Absolute’ in New Testament / Hellenistic Greek: A Proposal for Clearer Understanding.” *Journal of Greco-Roman Christianity and Judaism* 3:142–167.

Gorman, Vanessa B. 2020. “Dependency Treebanks of Ancient Greek Prose.” *Journal of Open Humanities Data* 6: 1. DOI: https://doi.org/10.5334/johd.13.

Gorman Vanessa B., and Robert J. Gorman. 2016. “Approaching questions of Text Reuse in Ancient Greek Using Computational Syntactic Stylometry.” *Open Linguistics: Treebanking and Ancient Languages*, eds. Giuseppe G.A. Celano and Gregory Crane. DOI: 10.1515/opli-2016-0026.

Hajič, J. 1998. “Building a Syntactically Annotated Corpus: The Prague Dependency Treebank.” *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevov*, 12–19. Prague.

Haug, Dag. 2015. “Treebanks in historical linguistic research.” In *Perspectives on Historical Syntax*, ed. Carlotta Viti, 188–202. Benjamins.

Haug, Dag T. T., Hanne M. Eckhoff, Marek Majer, and Eirik Welo.2009. “Breaking down and putting back together: analysis and synthesis of New Testament Greek.” *Journal of Greek Linguistics* 9:56–92.

Horrocks, Geoffrey. 2010. *Greek. A history of the language and its speakers*. 2nd ed. Chichester.

Jannaris, Antonius N. 1968 [1897]. *An historical Greek grammar chiefly of the Attic dialect as written and spoken from Classical Antiquity down to present time founded upon the ancient texts, inscriptions, papyri and present popular Greek*. Hildesheim.

Keersmaekers, Alek, and Mark Depauw. 2021. “Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri.” *Classics@* In this issue.

Keersmaekers, Alek, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. “Creating, Enriching and Valorizing Treebanks of Ancient Greek.” In

Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019), 109–117. Association for Computational Linguistics (ACL).

Korkiakangas, Timo. 2016. *Subject Case in the Latin of Tuscan Charters of the 8th and the 9th Centuries*. Commentationes Humanarum Litterarum 133. Helsinki.

Mambrini, Francesco and Marco Passarotti. 2012. "Will a parser overtake Achilles? First experiments on parsing the Ancient Greek Dependency Treebank." *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories (TLT11)*. Colibri 11.

———. 2013. "Non-projectivity in the Ancient Greek Dependency Treebank." *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, 177–186. Prague.

———. 2016. "Subject-Verb Agreement with Coordinated Subjects in Ancient Greek. A Treebank-based Study." *Journal of Greek Linguistics* 16:87–116.

Mayser, Edwin. 1934. *Grammatik der griechischen Papyri aus der Ptolemäerzeit mit Einschluß der gleichzeitigen Ostraka und der in Ägypten verfaßten Inschriften. Band II, 3. Satzlehre. Synthetischer Teil*. Berlin & Leipzig.

Morwood, James 2001. *Oxford Grammar of Classical Greek*. Oxford.

Thesleff, Holger. 1969. "Genitive absolute and Platonic style." *Arctos* 6:121–131.

Vierros, Marja, and Erik Henriksson. 2017. "Preprocessing Greek Papyri for Linguistic Annotation." In *Journal of Data Mining and Digital Humanities. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*, eds. Marco Büchler and Laurence Mellerin. http://jdmdh.episciences.org/paper/view/id/1385.

Vierros, Marja, and Erik Henriksson. 2021. "PapyGreek Treebanks: A Dataset of Linguistically Annotated Greek Documentary Papyri." *Journal of Open Humanities Data* 7: 26, pp. 1–6. DOI: https://doi. org/10.5334/johd.55.

Vierros, Marja 2018. “Linguistic Annotation of the Digital Papyrological Corpus: *Sematia*.” In *Digital Papyrology II. Case Studies on the Digital Edition of Ancient Greek Papyri*, ed. Nicola Reggiani, 105–118. Berlin/Boston. https://www.degruyter.com/viewbooktoc/product/486983.

Abbreviated references of papyrus texts

The Greek documentary papyri and links to the digital versions can be found by their abbreviations in the *Checklist of Editions of Greek, Latin, Demotic and Coptic Papyri, Ostraca and Tablets*. http://papyri.info/docs/checklist.

Footnotes

[back] **1.** This article was written with the project “Digital Grammar of Greek Documentary Papyri,” which received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 758481). https://www.helsinki.fi/en/researchgroups/digital-grammar-of-greek-documentary-papyri.

[back] **2.** Linguistic studies based on Ancient Greek treebanked data: Haug et al. 2009; Mambrini and Passarotti 2012; 2013; 2016; Celano 2014a; Gorman & Gorman 2016. For Latin, see e.g., Korkiakangas 2016.

[back] **3.** Morwood 2001: 140–141.

[back] **4.** Emde Boas et al. 2019: 624–625.

[back] **5.** One indicator of the rising interest that has grown slowly but steadily is the volume Evans and Obbink 2010.

[back] **6.** Thesleff 1969: 121.

[back] **7.** Thesleff 1969: 131.

[back] **8.** Jannaris 1897: 499.

[back] **9.** Mayser 1934 II:3: 66–74 (§157 B).

[back] **10.** See Soisalon-Soininen 1987: 177 for the Septuagint, and Fuller 2006.

[back] **11.** Soisalon-Soininen 1987: 177 and Mayser 1934 II:3: 68.

[back] **12.** Fuller 2006: 143–144 gives many examples from different grammars. The Cambridge Grammar mentions this as a second note of exceptions, saying that it can happen "infrequently" and usually when the GA precedes the matrix clause (Emde Boas et al. 2019: 625).

[back] **13.** We may also note that Soisalon-Soininen (1987) examined what type of Hebrew constructions were translated into Greek using GA in the Septuagint; often there was a certain type of clause (that could always be temporally translated) in the Hebrew text. He also admired the skill of those translators who were able to render the structure into Greek using the GA.

[back] **14.** Fuller 2006: 152.

[back] **15.** See e.g., Hajič 1998.

[back] **16.** Recently in Celano 2019.

[back] **17.** Bamman and Crane 2011.

[back] **18.** Landing page https://perseusdl.github.io/treebank_data/ or GitHub repository https://github.com/PerseusDL/treebank_data/tree/master/v2.1/Greek.

[back] **19.** Token counts for each author available below tables 2–4.

[back] **20.** http://sites.tufts.edu/perseids/. The names of the annotators of AGDT treebanks can be found in the GitHub repository's opening page (see note 18).

[back] **21.** Celano 2014b.

[back] **22.** Aesop, Apollodorus and Pseudo-Homer.

[back] **23.** See Gorman 2020 and https://github.com/vgorman1/Greek-Dependency-Trees.

[back] **24.** Bamman and Crane 2008.

[back] **25.** Digital Grammar of Greek Documentary Papyri, see above n. 1. The latest version of the corpus is available via https://papygreek.hum.helsinki.fi/. In this paper, we use the data of the stable release 1.01, published in July 6, 2021 (DOI: 10.5281/zenodo.5074307; Vierros and Henriksson 2021). The annotations of PapyGreek data go through a review process in which at least one person other than the original annotator checks and accepts the annotation (after possible corrections). The annotators of the PapyGreek/Sematia corpus are Arttu Alaranta, Iida Huitula, Sari Kock, Petri Lahtinen, Lauri Marjamäki, Jamie Vesterinen, Marja Vierros, and Polina Yordanova.

[back] **26.** See Vierros and Henriksson 2017, Vierros 2018 and Celano 2019.

[back] **27.** We follow the Guidelines of AGDT 2.0. (Celano 2014b) without the advanced syntactic/semantic layer but have also compiled additional guidelines for ambiguous cases and cases typical for papyrological documents. This is an evolving document with a link provided in the PapyGreek portal.

[back] **28.** The parts written by different hands are no longer cut to separate files, but the hand change is still coded within the document so that we can identify parts written by different people (acts of writing). The tokenization is similar to the one used in the Leuven project, in order to be able to make use of their automatically parsed files—see below on Duke-nlp, and Keersmaekers & Depauw (in this issue).

[back] **29.** Trismegistos Archive (later TM Arch) ID 256, mid-third century BCE.

[back] **30.** TM Arch ID 119, mid-second century BCE.

[back] **31.** TM Arch ID 26, late first century BCE–early first century CE.

[back] **32.** https://kiln.readthedocs.io/en/latest/.

[back] **33.** This approach is effective in finding standard GA constructions consisting of participle(s) and agent(s) in the genitive, but, unfortunately, cannot discover agentless GA's, as a query for a genitive participle marked with ADV returns too many results that cannot automatically be narrowed down to only contain the GA. In our experience, those cases are not common.

[back] **35.** Admittedly, Xenophon cannot be labeled merely a 'historian' as his works span many genres.

[back] **36.** The Constitution of the Athenians (also known as Old Oligarch).

[back] **37.** The existence of Herodotus' Book 1 in both corpora gives us a nice illustration on differences caused (possibly) by different tokenization, sentence division, and lastly annotation of certain constructions; a human factor is at play in some interpretations.

[back] **38.** Table 5 represents figures from the original corpus.

[back] **39.** It should be added that some of the petitions in the corpus are drafts and not all of them are preserved in their totality.

[back] **40.** One explanation for this can partly be the fact that many letters in PapyGreek corpus are short letters written on ostraca from a Roman garrison in Mons Claudianus, among which only one GA construction exists within the annotated letters: O.Claud.2.247.

[back] **41.** In contracts, though, the GA appears very often indicating the regnal year in the dating formula (e.g., βασιλευόντων Πτολεμαίου τ[ο]ῦ ἐπικαλουμένου Ἀλεξάνδρου καὶ Βερενίκης … ἔτους ις 'in the 16th year of Ptolemy, also called Alexander, and Berenike' P. Grenf.2.35). The formulaic parts of the language in papyri could be well-identified within further treebank-based studies, and it should also be relatively easy to treat them separately from the non-formulaic parts.

[back] **42.** AGDT, Aesop, sentence 10.

[back] **43.** PapyGreek, p.eleph.2dupl, sentence 10, unpublished.

[back] **44.** PapyGreek, p.enteux.26, sentence 3, unpublished.

[back] **45.** Translation from Bagnall & Derow, The Hellenistic Period (2004), Nr. 152.

[back] **46.** AGDT, Polybius, sentence 784.

[back] **47.** Translation published in Vol. I of the Loeb Classical Library edition,

1922–1927, taken from LacusCurtius website.

[back] **48.** See Haug et al. 2009; Haug 2015.

[back] **49.** Guidelines: Haug 2010 http://folk.uio.no/daghaug/syntactic_guidelines.pdf.

[back] **50.** PROIEL in GitHub: https://proiel.github.io/, https://github.com/proiel/proiel-treebank/. INESS: http://clarino.uib.no/iness/page.

[back] **51.** Queries performed by M. Vierros, May 13, 2020; query for type 1: #x:[pos!="C."] >adv #y: [morph="…p..g.*"] >sub #z: [morph="……g.*"]; for type 2: #x >adv #y: [morph="…p..g.*"] >sub #c: [pos="C."] >sub #z: [morph="……g.*"]; for type 3: #x:[pos="C."] >adv #y: [morph="…p..g.*"] & #x >sub #z: [morph="……g.*"]; for type 4: #x:[pos="C."] >adv #y:[morph="…p..g.*"] & #x >sub #z: [morph="……g.*"] & #x >sub #c: [pos="C."] (type4 query yielded no results).

[back] **52.** See Keersmaekers et al. 2019 and https://github.com/alekkeersmaekers/dendrosearch and https://github.com/perseids-publications/pedalion-trees/, many of which are based on automatically parsed text and checked by different people.

[back] **53.** Downloaded from the Duke Databank of Documentary papyri in 2016. See Keersmaekers and Depauw in this volume and https://github.com/alekkeersmaekers/duke-nlp.

[back] **54.** Search for type 1 GA was performed with the following constraints: [1] genitive participle verb; ADV relation (any suffix) + dependent node: [2] genitive; SBJ relation; type 2 GA with the constraints [1] genitive participle verb; ADV relation (any suffix) + [2] (depending on [1]) COORD relation + [3] (depending on [2]) genitive; SBJ relation (any suffix). As searches for types 3 and 4 did not yield many results, the automated parsing is most likely not handling that kind of coordination in quite the same way as AGDT, Gorman, and PapyGreek corpora; it is likely that some type 3 constructions exist within the type 1 query, and type 4 within the type 2 query. Alek Keersmaekers informed us that some technical issues concerning more than two coordinated words have since been resolved, but are not yet in the GitHub repository at the time of writing.

[back] **55.** E.g., P. Prag.1.34 (TM40943), Fr.B, 2–5: [ἀ]παιτήσεώς σοι γιγνομένης ἐξ ὑπαρχόντων μου πάντων καθάπερ ἐκ δίκης καὶ ἐπερ(ωτηθεὶς) ὡμολόγησα yielded three hits: 1) [SBJ [ἀ]παιτήσεώς] [ADV_CO γιγνομένης] 2) [ADV ὑπαρχόντων] [SBJ μου] and 3) [ADV ὑπαρχόντων] μου [SBJ πάντων]; the first one is a GA, but the two last present erroneous annotation, as μου and πάντων should be attributes, not subjects, of ὑπαρχόντων. Another example is P. Petr.2.26 (TM7625) where the phrase βασιλεύοντος Πτολεμ[αίου] τοῦ Πτολεμαίου Σωτῆ[ρος] has been counted twice due to the marking of both Πτολεμ[αίου] and Σωτῆ[ρος] with the label SBJ, even though Soter is the title of Ptolemaios; the same has happened in the exact same phrase in some other documents as well, but curiously not in all, like in P. Zen.Pestm.1 (TM1832), which has also yielded two hits, but is correct since the document really is a duplicate and contains the same phrase twice.

[back] **56.** The type 2 figures have been cut by half since the actual number of hits almost always included the same construction (at least) twice due to the count of both (or all) subjects separately; sometimes there have been more than two subjects, but cutting the number by half should provide a rough estimate of the actual number of constructions.

[back] **57.** With more time, it is, of course, possible to take samples and see the percentage of correct and false cases; according to Alek Keersmakers, the precision is high for genitive subject labeling and head attachment: over 90% when checked with 55 cases from the test corpus of papyri (current state of the art) including mostly letters that are easier to parse; the whole Duke-nlp in GitHub is not yet exactly at the same level. This gives us an important estimate for evaluating the results.

[back] **58.** And since all the data is available with academic licenses, it is always possible to further modify and annotate the data to fit your specific querying method, as we have done with the Kiln platform.