Preprocessing Greek Papyri for Linguistic Annotation

Marja Vierros, Erik Henriksson

Journal of Data Mining and Digital Humanities

Preprocessing Greek Papyri for Linguistic Annotation

Marja Vierros[1*], Erik Henriksson[2]

1, 2 University of Helsinki, Finland

*Corresponding author: Marja Vierros marja.vierros@helsinki.fi

Abstract
Greek documentary papyri form an important direct source for Ancient Greek. It has been exploited surprisingly little in Greek linguistics due to a lack of good tools for searching linguistic structures. This article presents a new tool and digital platform, "Sematia", which enables transforming the digital texts available in TEI EpiDoc XML format to a format which can be morphologically and syntactically annotated (treebanked), and where the user can add new metadata concerning the text type, writer and handwriting of each act of writing. An important aspect in this process is to take into account the original surviving writing vs. the standardization of language and supplements made by the editors. This is performed by creating two different layers of the same text. The platform is in its early development phase. Ongoing and future developments, such as tagging linguistic variation phenomena as well as queries performed within Sematia, are discussed at the end of the article.

keywords
Greek; papyri; linguistic annotation; treebank; dependency grammar; TEI EpiDoc XML; MySQL; Python; JavaScript

INTRODUCTION

Greek papyri from Egypt have preserved bigger and smaller entities of Greek as it was written by ancient speakers from ca. 300 BCE to 700 CE. There are different registers and styles found within a variety of different text types; the vernacular becomes visible in private letters and the official phraseology in contracts. Therefore, the papyrological corpus forms an important direct source for Greek linguists. The documentary papyrological corpus is freely available in digital form in the [Papyrological Navigator] (PN) platform, which also allows users to search both text strings and metadata (such as date and provenance). The search possibilities do not, however, easily yield to querying linguistic structures or variation in spelling or morphosyntax. Partly for this reason, the papyrological corpus has been left without much attention within the majority of linguistic research of Ancient Greek. A research project of author 1 ("SEMATIA: Linguistic Annotation of the Greek Documentary Papyri – Detecting and Determining Contact-Induced, Dialectal and Stylistic Variation" funded by the Academy of Finland) sought methods to make better use of the papyri for purposes of linguistic research. In this first phase we needed a way to preprocess the papyri into a form which could be linguistically annotated. The Sematia tool presented in this article results from this project but the tool is still being further developed. A new research project ["Act of the Scribe: Transmitting Linguistic Knowledge and Scribal Practices in Graeco-Roman Antiquity"] where author 1 is currently a researcher, is concentrating on scribes, their level of competence and their linguistic skills. We study the mechanisms of the language production in order to separate the technical effects from the linguistic and cognitive processes. This enables us to pinpoint the scribe's part in language change. We have added the possibility for implementing new metadata especially for the purposes of that project in Sematia. We approach the texts by dividing them by the "acts of writing" in order to distinguish each writer within one text. Sometimes a text is a product of one writer only, but

in many cases two or more different people have written in one document, attested by the change of handwriting.

I BACKGROUND

In this section, we will first briefly describe the digital papyrological corpus used in this project, as well as the nature of a papyrus text, in order to illustrate the basic requirements for preprocessing the data. Then, we summarize the linguistic annotation process in 1.2, essential for the later discussion on how we plan to utilize treebanks in this project. Lastly, in order to motivate the way in which we address the texts, we shortly discuss what we mean by linguistic variation in 1.3.

1.1 The Papyrological Corpus in Digital Form

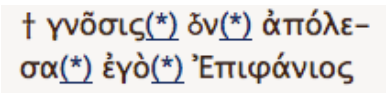
The platform Papyrological Navigator (PN) is the most important digital tool for papyrologists and anyone using papyri, potsherds and wooden tablets as primary sources for their studies of the Ancient World. It is an umbrella platform under which several databases with different scopes are linked together. Its history goes back to 1982, when a papyrological text corpus in digital form was formed at Packard Humanities Institute, resulting in a CD-ROM (PHI #7 Duke Databank of Documentary Papyri). A more detailed history is given in the information page at PN. At the moment the Duke Databank text corpus is open source and available online via the Papyrological Navigator and the texts have been migrated into [TEI EpiDoc XML] form. New publications are added to the corpus, old entries can be corrected and new data added via the Papyrological Editor by the papyrological community (the workflow is curated by an editorial team). Thus, the corpus is kept in an up-to-date, reliable state. Currently, it hosts ca. 70,000 Greek texts, 2,000 Latin texts and 1,000 Coptic texts. A word count is not available and texts vary from very short to extremely long. The PN also includes a search interface, where the texts, metadata, translations and images can be searched using different parameters.

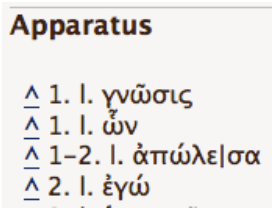
1.1.1 The Nature of a Papyrus Text and its realization in TEI EpiDoc XML

Papyri, like inscriptions, are seldom preserved in perfect condition. This results in gaps (*lacunae*) within the text. The ink may have faded in places, or the handwriting might be difficult to read, with the result that the editor cannot always be certain how to read each letter. Moreover, many texts contain a large number of abbreviations, because they come from the pens of professional scribes working with texts of an administrative nature. These features are marked in the paper editions according to the editorial conventions called the Leiden System, commonly agreed upon in 1931. For example, a *lacuna* is marked with square brackets, abbreviations are expanded within parentheses and uncertain letters have a dot under them. For a full list, see [Schubert 2009, 202–203]. The EpiDoc XML marks the same phenomena in TEI compatible tags within the text, e.g. <gap> for the *lacunae*, <uncertain> for the uncertain letters. The display in the PN shows the text in a traditional Leiden System layout (with the *apparatus criticus* below the text), but the text is stored in the GitHub repository in the XML form.

Example 1. The first two lines of P.Petra 1 6 in PN display layout (A) and in EpiDoc XML (B):

(A)





(B)

```
<lb n="1"/><g type="stauros"/> <choice><reg>γνῶσις</reg><orig>γνῶσις</orig></choice>
     <choice><reg>ὧν</reg><orig>ὃν</orig></choice> <choice><reg>ἀπώλε
<lb n="2" break="no"/>σα</reg><orig>ἀπόλε<lb n="2" break="no"/>σα</orig></choice>
     <choice><reg>ἐγώ</reg><orig>ἐγὼ</orig></choice> Ἐπιφάνιος
```

Although Example 1 exhibits no gaps or uncertain letters, it shows another feature that is highly relevant to our project and to linguists in general, namely, editorial corrections. Within the <choice> tag, the <orig> tag informs which form the ancient writer really wrote on the papyrus and <reg> what the editor thinks is the regular or standard form which was meant. A linguist is usually interested in the forms that the writer originally wrote, since they give us information on language change, phonology and the vernacular. However, with regard to our project it is highly important that the edited text contains the assumed standard forms, too. Using that information, the lemmatization and comparison between the original and standard forms are much easier to perform. Of course, we may be hesitant in several cases about what, in fact, is the standard we should be comparing with and if we agree with the editor's interpretation of what was sought after by the original writer. For discussions on this topic, see [Colvin 2009] and briefly [Vierros 2012, 25].

1.2 Treebanks

For Ancient Greek literature, two (constantly growing) linguistically annotated treebank corpora exist, as mentioned by [Haug 2014]: the Ancient Greek Dependency Treebank (currently ca. 558,000 tokens of Homer, Hesiod, tragedies) and the PROIEL treebank (currently ca. 230,000 tokens of the New Testament, Herodotus and later Greek), see also [Universal Dependencies]. These treebanks follow the Dependency Grammar originally used for Czech in the Prague Dependency Treebank outlined in [Hajič 1999]. The suitability of treebanks for historical linguistic research as well as dependency grammar for Ancient Greek has been recently discussed by [Haug 2015]. The most reasonable solution, in our opinion, was to follow the same framework of annotation also with the papyrological material. In this way we can utilize best practices and an annotation infrastructure in those projects as well as gain maximal synergy between the corpora of literary and documentary texts.

In the annotation process each word is supplied with a tag including its lemma, postag (i.e. string containing the part-of-speech and morphological analysis of the form), syntactic role and a reference to the head word. The analysis is performed according to the Guidelines for the annotation of Ancient Greek (see [Bamman and Crane 2008] and [Celano 2014] for versions 1.1 and 2.0, respectively). The annotation tool we have used is an editor called [Arethusa] in the [Perseids] platform. Arethusa first divides the text into sentences at certain punctuation (full stop, colon) and performs the tokenization, i.e. gives each sentence and each word within the sentence an ID number. It employs the [Morpheus] tool in providing each word with a lemma and with morphological analysis. This means that lemmatizing and morphological analysis are performed semi-automatically in the Arethusa editor; the human annotator must evaluate the correctness of the analyses where several options are possible in the case of homonyms and add forms in cases where the tool does not recognize the lemma (e.g. many Egyptian names in the papyri). The syntactic roles and dependencies have to be

analysed by the human annotator and implemented manually because a syntactic parser for Ancient Greek is still a desideratum; the first attempts have been reported by [Mambrini and Passarotti 2012].

Example 2. Treebanked sentence in XML format.

```
<sentence document_id="https://sematia.hum.helsinki.fi/edit/10" id="5" span="" subdoc="">
   <word form="Ἑρμίας" head="5" id="1" lemma="Ἑρμίας" postag="n-s---mn-" relation="SBJ" />
   <word form="ὁ" head="7" id="2" lemma="ὁ" postag="l-s---mn-" relation="ATR" />
   <word form="παρὰ" head="7" id="3" lemma="παρά" postag="r--------" relation="AuxP" />
   <word form="Πανίσκου" head="3" id="4" lemma="Πανίσκος" postag="n-s---mg-" relation="ATR" />
   <word form="κεχρηΑ" head="0" id="5" lemma="χρηματίζω" postag="v__r__---" relation="PRED" />
   <word form="." head="0" id="6" lemma="punc1" postag="u--------" relation="AuxK" />
   <word artificial="elliptic" form="αγορανομος" head="1" id="7" insertion_id="0001e" relation="ATR" />
</sentence>
```

The “postag” is a nine-place string marking each lemma with 1) part of speech 2) person 3) number 4) tense 5) mood 6) voice 7) gender 8) case 9) the degree of comparison, using certain agreed letters and numerals, e.g. “n” stands for nominative and “g” for genitive within the 8th place of the string, marking “case”.

1.3 Linguistic Variation

The documentary papyri include many different types of linguistic variation, which often cannot be found in the literary texts preserved via the manuscript tradition. Variation means the existence of competing linguistic forms either within one single speech community or a language as a whole. When we witness a change in a language, it is normally preceded by a great deal of synchronic variation, that is, many variants compete until one of them becomes popular and consistent. Studying the variants as such not only tells us a great deal about language change and the processes leading to it, but also about the community; where the people come from, and with whom they have interacted (contact induced variation). Some of the variants in papyri can be categorized as “scribal errors”, a category which is not always treated consistently. It may include mere slips of the pen, but sometimes even a difference of one letter may be an important phonological variant signalling changes in pronunciation. For example, the genitive singular of the word “wheat” (standard: πυροῦ) is written in two different nonstandard ways in the potsherds from Narmouthis (the potsherds, *ostraca*, are included in the papyrological corpus): πουροῦ (OGN I, 42 and 47) and ποιροῦ (OGN I 46 and 86). The latter (ποιροῦ) attests the merging of /y/ and /oi/ that was an internal development in Greek in the Roman period, but the former (πουροῦ) shows more the transfer of Egyptian, which did not have the front vowel /y/, and often the /u/ and /y/ were confused by Egyptians writing Greek, see [Dahlgren, 2016 and 2017].

In addition to spelling variants, we wish to present a couple of examples of morphosyntactic variation in order to make our treatment of the papyri more understandable. First, the phrase initial inflection strategy. Greek is an inflecting language where morphological case agreement is essential. Certain examples of case incongruence were earlier considered mainly “bad Greek”, but shown by [Vierros 2012] to present a pragmatic strategy for certain scribes; they only inflected the phrase initial words and left the rest of the words belonging to the same phrase in the nominative case. It also reflected the native language, Egyptian, of the writers, as it did not have case inflection. Also, the relative pronouns of the same writers were inflected according to the wrong head, thus evidencing contact-induced transfer from Egyptian.

A different type of dilemma is presented by some spellings that prevent us from making direct assumptions on what form the ancient writer aimed for. [Leiwo 2010] discusses, for example, how the phrase καλῶς ποιήσεις (a way of saying "please", "you do well…") is used; i.e. which form of a verb can act as its complement. Usually, an aorist participle complement denotes what is being asked. However, in the *ostraca* from Mons Claudianus, a form πέμψε is used (O. Claud. II 243, 2–3). In this particular case, it is difficult to say how it should be interpreted: straight up, πέμψε, would be the aorist indicative 3rd person singular form of the verb "to send" and this is how the automatic morphological tool would classify it. In the sentence it cannot be a 3rd person form since the phrase is directive. We could interpret it at least in two different ways. It could be an aorist imperative 2nd person singular, πέμψον, because unstressed /e/ and /o/ could be confused, especially by Egyptian native speakers, and the final /n/ could easily be dropped out. This is how the editors wish to regularize it. However, also the infinitive form, πέμψαι would be a phonologically possible interpretation here because the <αι> and <ε> are often confused in the papyri. All the forms discussed above were probably pronounced in the same way: /pémpsə/. The annotator may wish to mark up both options, the infinitive and the imperative, because the question here is whether the infinitive form was an accepted variant with this directive phrase or not.

II PREPROCESSING THE PAPYRI

In this section, we first present the idea of layering as a solution to preprocessing the papyrological data. Second, 2.2 contains the detailed description of how each XML tag is treated in the selection or deselection of elements for each layer. The technical side of building the platform and tool, for which author 2 was in charge, is described in 2.3.

2.1 Layers in Sematia

As mentioned in 1.1.1, the XML tags in the papyrus texts code important information. The tags are located inside the text and between words and letters. Similarly, the choices and apparatus entries for one word follow each other. In the treebank editor, a word is the basic element it tries to identify automatically. The EpiDoc XML texts cannot therefore be uploaded to the treebank editor as such, because the tags break up the words and the apparatus choices would all be included side by side if we only removed the tags. For the study of linguistic variation, we need first and foremost to know what the ancient author really wrote (and what is extant of what he wrote). However, the standard variant is useful to have for the sake of comparison. Moreover, the fragmentary nature of many texts makes the syntactic structure discontinuous, and therefore the editor's supplements may help in having a solid syntactic tree of a sentence, which is otherwise broken.

For these reasons, it seemed justified that we should create two different layers of the same text, each of which will be treebanked separately. First, the *original* layer contains only what has been preserved in the papyrus and in the form the ancient writer wrote them. For abbreviated words, for example, only the part that was written is taken into the original layer to prevent us annotating case inflection that the ancient writer did not produce. The *standard* layer, on the other hand, includes all the editorial work: the expanded abbreviations, supplements, as well as the standardized forms of misspelled words are all accepted. In this way, we get two different treebanks of one act of writing, and comparison can be made between them to see where the morphology differs.

Since treebanking does not allow us to mark all features relating to linguistic variation, we decided to add a third layer, where a new variation mark-up is added to the treebank XML. This very much concerns phonology and spelling, but can also benefit morphosyntactic

analyses. Moreover, different editors are not always consistent in what spellings they standardize. The variation layer is discussed in chapter IV (Future developments).

An important division of one document is performed before the layering. The change of handwriting, <handShift>, indicates a different person penning the letters. Thus, each act of writing gets its own layers and eventually treebanks. Also, the new metadata we enter (discussed in III), concerns each act of writing.

One caveat may be mentioned, although the present article is not the correct place to take the discussion very far. The original layer, in fact, contains some editorial work too, i.e., it does not present a so-called diplomatic transcript. The writing on the papyrus is usually without word divisions (in *scriptio continua*) and does not contain diacritical marks (accents, breathings, or *iota* subscripts). The word divisions and diacritics are part of the editor's interpretation and make the text readable. We have not moved towards a diplomatic transcript in the original layer for the sake of readability as well as to facilitate the automatic lemmatization and morphological analysis. If the annotator disagrees with some word divisions or diacritics, s/he has the possibility to make a change in the text in the Arethusa tool. However, in that case the interpretation should be well supported and the same correction should be suggested to the Papyrological Navigator.

2.2 How tags were treated

This chapter consists of a full discussion of how the TEI EpiDoc XML tags are treated when creating the original vs. the standard layer (for a quick glance, the same information is collected in Table 1 at the end of this section). It was important to keep the word count, i.e., keep the tokenization the same in both layers, so that the word-for-word comparison is possible between the layers by using the word-IDs. We use "dummy" elements to replace the parts not included in the layers on account of tokenization. Another reason for using dummy elements is to help the annotator to notice the missing parts of the text. The annotator will clearly see that something is missing either between the words or at the end of an abbreviation when s/he sees the dummy element. For this reason, the dummy element is written in capital letters.

2.2.1 Editorial corrections: <choice>, <reg>, <orig>, <corr> and <sic>

The element <choice> usually contains two alternatives. First, <reg> gives the standardized, regularized version, and is thus selected for the standard layer. On the other hand, <orig> consists of what was originally written on the papyrus, and is naturally elected for the original layer. E.g. from

<choice><reg>γνῶσις</reg><orig>γνο͂σις</orig></choice>

we choose γνῶσις for the standard layer and γνο͂σις for the original layer. Sometimes the editor may have suggested two different possibilities for regularizations, or another scholar may have suggested a new interpretation. In those cases, the platform allows the user to choose one of the options to the text which will be annotated (see below 2.3.3).

Pure scribal mistakes are sometimes coded with the pair <corr> and <sic>. Then, from e.g.

<choice><corr>τιμὴν</corr><sic>τμμὴν</sic></choice>

ISSN 2416-5999, an open-access journal

we choose what is marked corrected with <corr> for the standard layer, i.e. τιμὴν, and what is marked with <sic> for the original layer, i.e. τμμὴν.

2.2.2 Abbreviations: <expan>, <ex>
Words are abbreviated in different ways in the papyri. Sometimes only the end of the word is left unwritten (and it usually has some sort of abbreviation mark at the break up point). In TEI EpiDoc XML, the <expan> tag surrounds the whole word which is abbreviated in its expanded form and, within the <expan> tag, the part which was left unwritten is surrounded by the <ex> tag. For example, when the word στερεοῦ is abbreviated by leaving out the ending οῦ, it is written στερε(οῦ) according to the Leiden System, but in EpiDoc XML it is marked:

<expan>στερε<ex>οῦ</ex></expan>

In this case, we take the whole word in expanded form into the standard layer (στερεοῦ) and for the original layer we choose only what was written on the papyrus, i.e. στερε, now added with the dummy for abbreviation: A. Thus in the original layer we get στερεA. The annotator now immediately sees that the scribe has not written the ending of the word, and can annotate the word for lemma and other factors that are visible, but not, in this instance, by its morphological case.

Some words have been abbreviated only with a certain abbreviation mark. One of the most common is the sign L for ἔτος, "year". In this case the word is most often opened up in the genitive and marked within the parentheses in the Leiden System: (ἔτους). The markup is:

<expan><ex>ἔτους</ex></expan>

The whole word in expanded form, ἔτους, is chosen for the standard layer and for the original layer it is substituted with the marker A. The annotator may be confident enough to lemmatize the word for ἔτος, but otherwise the morphological analysis should be left open.

2.2.3 Supplements and omissions: <supplied>, <surplus>
When there is a hole in the papyrus, it may be possible for the editor to make an educated assessment about what probably was written in the gap and restore it. Especially if the gap was short (only a few letters) or if the missing part is in a formulaic part of a text, the parallel documents help in restoring the text. When text is restored in the *lacuna*, it is written inside square brackets in the Leiden System, and in TEI EpiDoc XML it is marked with the tag <supplied> with the reason attribute "lost". The markup can go over word boundaries. For example:

μ[ε]λίχρως = μ<supplied reason="lost">ε</supplied>λίχρως
ὄντ[ος ἐ]ν = ὄντ<supplied reason="lost">ος ἐ</supplied>ν

We choose the restorations for the standard layer without brackets, that is, we get μελίχρως and in the latter example two words: ὄντος ἐν. This way, the linguistic annotation tool correctly recognizes these words. For the original layer, however, the supplements are not taken in, since we cannot be sure if the editor has been right; the ancient writer could have written a nonstandard variant even in a short space. The supplement receives the dummy marker SU in the original layer: μSUλίχρως and, in the case of two words, both get their own marker: ὄντSU SUν. Especially when there are several words in a *lacuna*, it is important that

each word (and punctuation mark) is counted in the same way in both layers in order to keep the tokenization the same.

Another type of supplement is when the editor of the papyrus thinks that the ancient writer has not, by mistake, written something we would expect. The editor can add what was omitted using angle brackets in the Leiden System; in EpiDoc XML it is rendered with the supplied tag with the reason attribute "omitted":

ἀπ<ε>γραψάμην = ἀπ<supplied reason="omitted">ε</supplied>γραψάμην

Again, we choose the supplement for the standard layer as the editor suggests: ἀπεγραψάμην. For the original layer the supplement is replaced with the dummy marker OM, i.e. ἀπOMγραψάμην.

The opposite case is <surplus>, which indicates text which the original writer wrote, but the editor considers superfluous. This surplus text is replaced with the marker SR in the standard layer but included as such in the original layer.

2.2.4 No supplements in lacuna*: <gap>*
When there is a *lacuna* in the papyrus in which the editor has not been able to suggest a supplement, this <gap> is replaced with the dummy element G both in the standard and in the original layers. The reason is that, also when annotating the standard layer, the annotator should see if the sentence is not whole.

2.2.5 Uncertain letters: <unclear>
The 'conscience' of a papyrologist, the underdot, signals that a letter is only partially preserved or so faded that the editor cannot be certain beyond a doubt which letter the ancient writer wrote. He makes an assumption based on the ink traces he sees, writes the letter he assumes has been written in the papyrus, but puts a dot under the letter in the edition. In EpiDoc XML those letters are marked with the tag <unclear>:

Ἀλεξάνδρο̣υ = Ἀλεξάνδρ<unclear>ο</unclear>υ

In the standard layer it was an easy decision to include the uncertain letters in the same way as the supplemented letters. However, it was difficult to decide how to address the problem in the original layer, since we need the letter without markers interfering with the word recognition in the annotating environment. We decided to take the uncertain letters into the original layer in the same way as into the standard one. This may result in sometimes annotating a word which will later be read as another word. However, that may happen even in cases where the editor has not used underdots. Moreover, the annotator need not annotate the word at all if s/he does not trust the reading. The annotator has the possibility to change the text in the annotating framework, as mentioned previously in 2.1.

2.2.6 The apparatus: <app>
In the same way as above with <choice> (2.2.1), the *apparatus criticus* entries can include several options on what the editor or other scholars suggest for the readings. Tags are, e.g., <app type="alternative"> or <app type="editorial">. We have again decided to give the power of decision to the user; s/he can choose the best alternative to be included in the text which will be uploaded to the annotation tool.

	Original layer	Standard layer
<choice>: <reg>/<orig>	Text within *<orig>*	Text within *<reg>*
<choice>: <corr>/<sic>	Text within *<sic>*	Text within *<corr>*
<expan>	Text within *<expan>*	Text within *<expan>*
<ex>	dummy element: A	Text within *<ex>*
<supplied reason="lost">	dummy element: SU	Text within the tag
<supplied reason="omitted">	dummy element: OM	Text within the tag
<surplus>	Text within the tag	dummy element: SR
<gap>	dummy element: G	dummy element: G
<unclear>	Text within *<unclear>*	Text within *<unclear>*
<app>	user in Sematia chooses	user in Sematia chooses

Table 1. The treatment of TEI EpiDoc XML tags in the original vs. standard layer in Sematia.

2.3 Technical realisation

In this section, we describe the technical realization of "Sematia" as a web-based tool for creating, managing and querying the *original* and *standard* layers of Epidoc XML texts. We begin by sketching the overall data structure of Sematia, and go on to discuss how the system automatically extracts metadata and creates the two layers from imported documents. We then describe Sematia's integration with the Perseids API, and finally, how the annotated layers can be queried in Sematia.

Sematia is hosted on a University of Helsinki server at https://sematia.hum.helsinki.fi and is publicly available to everyone with a Google account, which is required for logging in. Alternatively, the tool can also be installed locally (but without the database) from the open source code available at https://github.com/ezhenrik/sematia/. The back-end of Sematia was developed with Python and MySQL, and the client-side interface with HTML and JavaScript.

2.3.1 Database

From the perspective of scribal production of papyri, the <handShift> elements in EpiDoc XML files are crucial, as they divide the document into parts penned by different persons (see also section 2.1 above). In Sematia, these "hands" each get their own linguistic layers (*original* and *standard*), as well as metadata (discussed in detail in III). Moreover, the documents imported to Sematia can be described with metadata about composition date and provenance. This results in the following database schema (fields are in parentheses):

- Document (id, user_id, XML, HTML, date, provenance)
- Hand (id, document_id, no, [metadata fields])
- Layer (id, hand_id, type, treebankXML, settings)
- User (id, name)
- User Document (id, user_id, document_id)

In this schema, each "layer" record is linked to a single "hand", which, in turn, is a child of a "document" record that belongs to the "user" who imported the document to Sematia. The document table contains fields for the source XML, the HTML-conversion (see 2.3.2 below) as well as date and provenance metadata. To avoid duplicating any data, no actual text is stored in the hand table; its only purpose is to serve as metadata storage for each act of writing. Lastly, the layer table has fields for the layer type, treebank annotation XML (see 2.3.5) and user settings for manually chosen variants (see 2.3.4, item 6).

2.3.2 Importing documents

In order to minimize the effort to create the layers in Sematia, we have automatized the workflow wherever possible. Thus, when the user imports a document (by entering the document URI into the system), Sematia first parses the XML and calculates the number of <handShift> elements in the document, which mark the boundaries of different acts of writing. A corresponding number of hand records is then created, as well as the *standard* and *original* layer records for each hand. Initially, these records are created as empty templates, to be filled in the stages described below.

Since the actual layering happens within the browser using JavaScript code (see 2.3.4 below), the XML tree is next converted to an HTML string that can be manipulated using the Document Object Model (DOM) interface. The following template is used in the conversion:

> Element nodes: `<span class=[tag name] data-[attribute key]=[attribute value]>`
> Text nodes: `<span class="text">[text]</span>`

The converted HTML string is then saved to the database.

2.3.3 Document metadata

Next, Sematia tries to populate the metadata fields of the new document record automatically via PN's Apache Solr API available at http://papyri.info/solr/select/?q=id:[document id]. At the time of writing this, Sematia is configured to fetch date and provenance metadata from this public API, in case these data are available for the imported document. As regards PN's date metadata, Sematia includes a mapper which converts the diverse formattings (e.g. "II spc", "II/IIIspc", AD709") to a machine-readable form ("101-200", "101-300" and "709-709", respectively). There is also an interface in Sematia for editing the metadata fields manually.

2.3.4 Creating the layers

The layering process is described in the following steps. The layer is created client-side in the browser from the HTML conversion of the imported document, using the jQuery javascript library.

1. The user chooses the layer and hand she wishes to work on, for example, *standard* layer of the 1st "hand" of Petra 1.1.
2. On a new page, the HTML conversion of the document (e.g. Petra 1.1) that contains the selected hand is loaded into the DOM tree.
3. The elements outside the selected hand (e.g. elements within hand 2 and 3, if hand 1 was selected) are hidden and marked for exclusion from layering. The motivation for loading the whole document first and hiding irrelevant sections later is the fact that <handShift> elements in EpiDoc XML documents may appear on different levels of the XML hierarchy. Due to this discrepancy, loading only the elements between two <handShift> tags would risk creating invalid structures in the DOM.
4. General formatting is applied, e.g. <lb> elements (line break) that have the attribute break="no" are removed in order to prevent unintended word breaks. Some CSS-styling is also applied in order to highlight different elements to the user (see item 6).
5. The layer is enabled according to the rules discussed in 2.2, by marking each element with a data-attribute either for exclusion or with the replaced value. For example, if the layer type is *original*, to each <ex> element is added an attribute with the segment

"A". The <supplied> element is a special case as it may contain several words as well as punctuation marks. In the *original* layer, we want to replace each word and punctuation with "SU" or "OM" (depending on the value of the "reason" attribute) to maintain the same word count in all layers. Likewise, we had to make sure that the tokenization would work the same way in both Sematia's layering tool and Arethusa's treebanking service. For these reasons, the regular expressions used to split up words in Sematia follow Arethusa's tokenization rules as closely as possible. For example, Arethusa has been configured to deal with crasis (e.g. κἀγώ, "I too") by treating the merged words as separate. In Sematia, a similar mechanism is currently under development.

6. In some cases, the editor of the papyrus has provided multiple readings for the same text part, contained in <choice> or <app> elements in the XML. Only one of the readings should end up in the layer, making it necessary for the user to choose the preferred interpretation manually. This feature was implemented by adding a click event listener to the elements that may have multiple readings, which allows the user to make the choice simply by clicking on the preferred variant. The manual edits are saved to the database and automatically enabled whenever the user returns to view or edit the layer.
7. Finally, the layer is created by collecting the new values in the act of writing that the user is working on. The resulting text is loaded into a panel next to the HTML version.

2.3.5 Perseids API integration

Sematia uses the [Perseids] Project API (https://sosol.perseids.org/sosol/api/v1) to handle the treebank annotation of the layers. We opted for a strong integration with the Perseids Project, since it is home to the syntactic annotation framework we use, [Arethusa]. Moreover, the Perseids platform offers a centralized review process for the annotations, which helps us to control the quality of the treebanks uploaded to Sematia.

The integration works out roughly as follows: First, a layer is created in Sematia according to the steps described in the previous section. Next, Sematia prepares the layer into a treebank annotation template using Perseids' tokenization and transformation tools (https://github.com/perseids-project/perseids-client-apps), which is then POSTed to Perseids for annotation. When the annotation is finished, it is placed in a review queue, where it can be approved, sent back to the annotator for revision or rejected by one of Sematia's administrators. Finally, after the approval, Perseids sends the treebank back to Sematia via a public GitHub repository dedicated to Sematia's finalized treebank annotations (https://github.com/ezhenrik/sematia-tb).

2.3.6 Queries

Sematia also includes a preliminary set of tools (https://sematia.hum.helsinki.fi/tools) for exporting the treebanked layers as a single .zip archive, listing frequencies of tokens in the treebanks, visualizing the data as a hierarchical document cluster, as well as for searching for occurrences of morpho-syntactic features or text segments in the treebanked layers. Using the search functionality, users can limit the results with metadata filters (e.g. document date) and combine them with regular expressions targeting individual fields of the treebanks (e.g. syntactic relation), which makes it possible to create highly specific queries on the data. For example, if one wishes to find instances where a participial verb form acts as the subject, the Relation field is filled as `SBJ` and the Postag field as `^....p`. The search tools are currently in a very preliminary stage, but the future development of Sematia will be focused on extending this particular area.

III METADATA

3.1. Metadata in existing databases

The metadata which concern the actual papyrus document can be found via the Papyrological Navigator from several different databases, e.g the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV) has collected information on the date and provenance of the text, the original title and the subject matter (in German); similarly, the Trismegistos portal adds the metadata of people involved and places mentioned, to mention a few aspects. For the needs of the project "Act of the Scribe" we wish to add metadata that would help in the identification of the writers as well as the linguistic register. In addition to that, the date and provenance is extracted automatically for each document from the PN, as discussed in 2.3.2.

3.2. Metadata to be added

The new metadata always concern one act of writing; that is, all writers in one papyrus get their own metadata field. It is divided into four sections: Handwriting, Writer and author, Text type, and Addressee.

3.2.1 Handwriting

The printed editions of papyri quite often have some sort of description of the handwriting, at least for the main hand of the text. Moreover, later research may have identified the hand as the same as in some other text, or made some other observations on it. However, if the current user of Sematia has seen the original text or a photograph of it, s/he can add his/her own custom evaluation to the handwriting. We included four subfields for describing handwriting. The first two, "Description in the edition" and "Custom description", are free text fields serving mainly the user as a reference. The third field is a drop-down list for the level of professionalism with four possibilities to choose from: Not known, Professional, Non-professional and Practised letterhand. The first is applicable when there is no description or a photograph or possibility to check the original. The last option is something between the professional and non-professional; a person who is accustomed to writing, but has obviously not received scribal training. The fourth subfield is reserved for entering a list of texts where the same handwriting is found. This list is stored as a JSON string in the database and may be used in the future for connecting the acts of writing by the same person in queries.

3.2.2 Writer and author

In our project, we are interested in distinguishing the linguistic acts of the actual writer (usually a scribe who has received more or less education) from those of the author of the text, who may have dictated the text or given written and/or oral instructions. Moreover, in official contracts there may be a scribal official 'responsible' for the text, e.g. a notary who may even sign the document with his own name, but is not the actual writer of the document, like the *agoranomoi* from Pathyris discussed by [Vierros 2012]. For these reasons, we have three categories which can be filled in, if the information is available, but left blank, if not: "Actual writer", "Scribal official" and "Author". For each one, there are three fields to be filled in: Name, Title and [Trismegistos] Person ID number. Later, when the corpus has a sufficient amount of texts, this information can be used, for example, for connecting people with similar titles to the similar use of language, or even finding texts that have been written/authored by the same person.

3.2.3 Text type and Addressee

The genre of the text naturally has an influence on the language used. A private letter belongs to a different register than a notarial contract. The addressee has a similar impact. The text is more formal if written to a superior than if written to a peer or subordinate. Therefore, it is important to gather this metadata when possible. We have added a drop-down list for the text type trying to cover the basic text types found in the papyri but also limiting the list to quite general categories (e.g. "contract" with certain subfields, "letter" with certain subfields, among others). For the addressee, we wanted a general description selected from a drop-down list: "official", "private" or "not known/applicable". The first two options get subfields with the subfields "superordinate", "peer" and "subordinate". In addition, there are fields for the addressee's name, title and Trismegistos Person ID number.

IV ONGOING AND FUTURE DEVELOPMENTS

4.1 Variation layer

Research on linguistic variation, discussed above in 1.3, is the driving force on building the Sematia corpus. Quite a number of such phenomena can be queried by comparing the original and standard layers. For example, if we are interested in morphological case agreement, the standard layer includes the grammatically 'correct' versions and the original has the variant forms. A search comparing, e.g., the case coding included in the postag of each word, reveals when a word has been written in an unexpected case (and similar comparisons can be made for mood, person, tempus, etc.). The biggest missing block of linguistic information concerns phonology, since spelling is not taken into account in the existing Treebank templates. This issue is to some extent addressed in the new database of Text Irregularities within the [Trismegistos] platform compiled by [Depauw and Stolk 2015]. Their data concerns the whole Duke Databank of Documentary Papyri and is collected phoneme by phoneme and based on the editorial corrections (i.e. the tags within the <choice> element, cf. 2.2.1). However, the editorial corrections are not always present in the DDbDP (for example certain editors have not necessarily thought it worthwhile to regularize all confusions of spellings between ι and ει) and it is not always clear if some confusion by a writer concerned the phoneme or the morpheme, i.e. whether the variation had phonological or morphological basis. For these occasions and for the greater accuracy in studying the linguistic variation, we plan to add a variation layer in Sematia. The treebank XML of the original layer would be duplicated and a new variation tag added for those words where variation exists.

The variation tagset in all its depth is still under consideration. We could have a tag for variation, <var>, and define it with different type attributes for phonology, morphology, and syntax. The types could be further defined with different values, e.g. for the immediate context, if that seems to play a role (in phonology at least). Also, a certain variation could be defined with two or more options, for example suggesting that we are fairly certain that a feature is either phonological confusion (e.g. of αι instead of ε) or a morphological one (e.g. confusion of aorist infinitive or imperative endings) or both at the same time.

4.2 Queries

Several tools for querying treebanked data already exist. Both Ancient Greek Treebank corpora can be queried with, e.g. [SETS Treebank Search], [PML Tree Query Engine] or [XQuery/BaseX] (see also [Universal Dependencies]). Moreover, the PROIEL corpus is available in INESS query interface. They employ somewhat different query languages, but all support detailed and complicated linguistic queries from the treebanked data. As mentioned in 2.3.6, all the available treebanked data can be exported, either all layers as one .zip archive or

the original layers and the standard layers separated as their own sets. Some quering possibilities have already been integrated in the platform itself (see 2.3.6), but they are still in a testing and developing stage. The important feature is to allow comparative queries between the original and standard layers. For example, one can search for instances where the Original layer has a dative case (Postag field: `^.......d`), but the Standard layer has a genitive case (Postag field: `^.......g`). The searches can also be performed on or limited by our new metadata.

Conclusion

In this article, we have described a process in which individual texts from the corpus of documentary Greek papyri can be preprocessed for the purposes of linguistic annotation. The annotation follows the same framework as other corpora of Ancient Greek texts. For the first time we can automatically separate the original text written by the ancient writer from the editorial interpretation. The original layer can be studied in its own right as well as compared with the standardized version. We have not disregarded the results of the hard editorial work devoted to these texts in the previous centuries, as they form the parallel layer of the text. The layers enable the comparison of linguistic variants abundant in the papyri to the scholarly standard forms. The tool is currently optimized for retrieving the texts from the Papyrological Navigator, but there is no impediment to modify it to be used for other texts which are encoded in EpiDoc XML, such as many epigraphic corpora.

References

For the abbreviations of papyrological editions, see Checklist of Editions of Greek, Latin, Demotic, and Coptic Papyri, Ostraca, and Tablets, of which the updated version is found online: http://papyri.info/docs/checklist.

"Act of the Scribe: Transmitting Linguistic Knowledge and Scribal Practices in Graeco-Roman Antiquity" http://blogs.helsinki.fi/actofscribe/.

Arethusa: available via Perseids sign in: http://sosol.perseids.org/sosol/signin.

Bamman, D. and Crane, G. Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank (1.1). The Perseus Project, Tufts University 2008. http://nlp.perseus.tufts.edu/syntax/treebank/greekguidelines.pdf.

Bamman, D. and Crane, G. The Ancient Greek and Latin Dependency Treebanks. *Language Technology for Cultural Heritage*, ser. Foundations of Human Language Processing and Technology. Springer (Berlin–Heidelberg), 2011:79–98.

Bamman, D., Mambrini, F. and Crane, G. An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. *Proceedings of the 8th Workshop on Treebanks and Linguistic Theories (TLT8).* 2009;8. http://www.perseus.tufts.edu/~ababeu/tlt8.pdf.

Celano, Giuseppe G. A. Guidelines for the annotation of the Ancient Greek Dependency Treebank 2.0. 2014. https://github.com/PerseusDL/treebank_data/edit/master/AGDT2/guidelines

Colvin, S. The Greek Koine and the Logic of a Standard Language. *Standard Languages and Language Standards: Greek, Past and Present*. Ashgate (Farnham), 2009:33-45.

Dahlgren, S. *Outcome of long-term language contact: Transfer of Egyptian phonological features onto Greek in Graeco-Roman Egypt.* University of Helsinki, doctoral dissertation. 2017. http://urn.fi/URN:ISBN:ISBN 978-951-51-3218-5.

Dahlgren, S. Towards a definition of an Egyptian Greek variety. *Papers in Historical Phonology.* 2016;1: 90–108. http://journals.ed.ac.uk/pihph/article/view/1695.

Depauw, M. and Stolk, J. Linguistic Variation in Greek Papyri: Towards a New Tool for Quantitative Study. *Greek, Roman, and Byzantine Studies*. 2015;55:196-220.

Hajič, J. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevov.* Charles University Press (Prague), 1998:12-19.

Haug, D.T.T. Computational Linguistics and Greek. *Encyclopedia of Ancient Greek Language and Linguistics.* Brill Online, 2014 (First appeared 2013; last online update November 2013).

Haug. D.T.T. Treebanks in historical linguistic research. *Perspectives on Historical Syntax*. John Benjamins, 2015:188-202.

Haug, D.T.T., Eckhoff, H. M., Majer, M., Welo, E. Breaking down and putting back together: analysis and synthesis of New Testament Greek. *Journal of Greek Linguistics.* 2009;9:56-92.

INESS (Norwegian *Infrastructure for the Exploration of Syntax and Semantics*): http://iness.uib.no.

Leiwo, M. Imperatives and other directives in the Greek letters from Mons Claudianus. *The Language of the Papyri.* Oxford University Press (Oxford), 2010:97-119.

Mambrini, F. and Passarotti, M. 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 2012;11.

Morpheus: https://wiki.digitalclassicist.org/Morpheus.

nltk http://www.nltk.org/.
Papyrological Navigator: http://papyri.info/.
Perseids: http://sites.tufts.edu/perseids/.
PML Tree-Query Engine: http://lindat.mff.cuni.cz/services/pmltq/#!/home.
Schubert, P. Editing a Papyrus. *The Oxford Handbook of Papyrology.* Oxford University Press (New York), 2009:195-215.
Sematia: http://sematia.hum.helsinki.fi/.
SETS Treebank Search: http://bionlp-www.utu.fi/dep_search
TEI EpiDoc XML: http://sourceforge.net/p/epidoc/wiki/Home/.
Trismegistos Portal: http://www.trismegistos.org/.
Trimegistos Text Irregularities: http://www.trismegistos.org/textirregularities/
Universal Dependencies: http://universaldependencies.org/.
Vierros, M. *Bilingual Notaries in Hellenistic Egypt. A Study of Greek as a Second Language.* KVAB (Brussel), 2012.
[XQuery / BaseX] http://docs.basex.org/wiki/Startup.