Linguistic Markup and Dialectal Variants. The Perspective of the Digital Corpus Supplementum Epigraphicum Creticum (e-SEC)

Alcorac Alonso Déniz

Chapter 12

Linguistic Markup and Dialectal Variants. The Perspective of the Digital Corpus *Supplementum Epigraphicum Creticum* (e-SEC)

Alcorac Alonso Déniz

Abstract: In recent years, several models for the linguistic markup of digital Ancient Greek and Latin corpora have been developed. Thanks to these developments, the once-challenging task of annotating and extracting the linguistic features of digital texts has been considerably simplified. This is, in particular, owing to various lemmatisers and annotators. In this chapter, I will address the challenges faced by the linguistic markup of corpora of texts written in non-standard Ancient Greek, particularly in dialects other than Classical Attic. In the first sections, I review the existing literature on the markup of linguistic variations in documentary papyri from Egypt. Secondly, I address the specific issues encountered in the process of linguistic markup vis-à-vis the dialectal corpus of Ancient Greek inscriptions from Crete, which, owing to their particularities, constitute a useful testing ground for the solutions hitherto envisioned. Finally, I propound an alternative: a 'lightweight linguistic markup' based largely on the guidelines of the Text Encoding Initiative (TEI).

Keywords: Linguistic variants, Greek dialects, Markup, TEI, Crete, Epigraphy.

Introduction

An increasing number of corpora of Ancient Greek texts (including inscriptions) have been, or are being, edited in XML format with a standard annotation following the EpiDoc Guidelines, derived from the standard of TEI.[1] At the same time, efforts to develop linguistic annotation tools for Greek and Latin documents have proliferated during the previous decade. In particular, lemmatisers and annotators have facilitated the work of texts' linguistic description, having helped to create annotated corpora bearing linguistic information.

This chapter is part of an ongoing project with two purposes: 1) the development of a digital epigraphic supplement of Ancient Greek inscriptions from Crete (e-SEC); and 2) the creation of a linguistic annotated corpus of Ancient Greek inscriptions from Crete, whose development could be deployed to better describe the history of Greek in Crete from the first written records until the 4th century AD (GIC or *Grammar of Cretan Inscriptions*).

A two-fold project: e-SEC and GIC

The main objective of this project (e-SEC) is to develop a digital edition of a supplement of the alphabetical Greek inscriptions from Crete, which are not included in the four volumes of Guarducci,[2] from the first documents attested in the late 8th century BC, to the beginning of the early Byzantine period (mid 4th century AD). The e-SEC project will include the texts written on all kinds of supports and will produce the philological editing of each text, followed by translations and a commentary, accompanied by photographs of the respective artefacts. The e-SEC project forms part of an ongoing effort on the part of the Histoire et Sources des Mondes Antiques (HiSoMA) laboratory in Lyon (UMR 5189), members of which uphold a deeply-rooted tradition of epigraphic projects in various areas of the Greek world, including Boeotia, Thessaly, Thasos, Syria, and Jordan.

While e-SEC is still in its nascency, a relational database (FileMaker Pro) has already been created containing information on each document. Furthermore, three epigraphical campaigns (2017, 2018 and 2019), funded by the laboratory HiSoMA, have been performed across several Cretan archaeological collections. The data collected in these campaigns (which also include photographs and squeezes) forms the core of this database.

The second part of the research project involves a *Grammar of Cretan Inscriptions* (or GIC, following the French *Grammaire des inscriptions crétoises*). This facet of the project aims to sketch and describe the history of the Greek language in Crete. It does so on the basis of the linguistic analysis of the epigraphical documents found on the island from the early 8th century BC to the mid 4th century AD. To this end, the project has foreseen the linguistic annotation of the nearly 3000 Cretan Greek inscriptions (viz. those of the e-SEC corpus, as well as those of Guarducci between 1935 and 1950). The ultimate goal of this markup is the integration of

[1] All the links referenced in this paper were accessed on 16 May 2020.
[2] Guarducci 1935-1950.

all the linguistic phenomena studied under the various rubrics of Ancient Greek grammar. To this end, special attention ought to be paid to the different regional varieties of Greek historically in evidence on the island of Crete. This project moreover concerns documents inscribed in Doric and Attic-Ionic Koine.[3]

Linguistic context of Ancient Crete

Ancient Crete stands out as a particularly interesting epigraphic and linguistic melting pot from the Bronze Age period down to the Classical, Hellenistic and Roman Ages. The first written documents attested on the island are the hieroglyphic inscriptions dated *c.* 2100-1700 BC and, immediately thereafter, the syllabary script conventionally known as 'Linear A' (1850-1450 BC). Both scripts remain undeciphered,[4] as well as the written system of the famous Phaistos Disk and the double-headed axe of Arkalokhori.

The first texts in Greek occur on clay tablets written in Linear B syllabary (*c.* 1450-1190 BC). Following this period, Greek inscriptions (including laws, decrees, letters, dedications, epitaphs in both prose and verse, etc.) are composed using different, local versions of the Greek alphabet from *c.* 700 BC, all of which get usurped by the Ionic alphabet between *c.* 400-350 BC. Interestingly, the Greek alphabet was also used for another undeciphered language, conventionally known as 'Eteocretan', as attested to on the east of the island (Dreros and Praisos; 8th-3rd century BC). Finally, the first documents in Latin occur in the 1st century BC following the installation of the Roman colonies in *Cnossus*, *Lictus* and *Cydonia*, following the island's conquest by Quintus Caecilius Metellus in 67-66 BC.

As for the Greek of the first millennium, two periods can be established from a dialectal point of view. These regard the variety of the language(s) attested in the inscriptions:

- Until *c.* 5th century BC the Doric dialect predominates, although it is far from being unitary.[5] In particular, the dialect of Cydonia in the east differs greatly from the dialect attested in the central and eastern-central areas.
- In the 3rd century BC,[6] many inscriptions were still written in Doric. We find, however, various degrees of influence of the Attic-Ionic Koine. In the same period, documents written in Doric Greek exhibit some peculiar features, which differ from region to region. Like the Greek of contemporary Crete, the Greek of Ancient Crete was far from being of a unitary kind.
- The Doric dialect peters out from the 1st century BC.[7]

Utility of the analysis of variants for epigraphy

The importance of studying the varieties of a regional corpus is, for linguists, immeasurable. The analysis of linguistic variation (on different grammatical levels), as found in written records from the past, performs an essential task within historical linguistic research.

Moreover, the study of dialects can help to corroborate the chronology and geographical origins of an inscription, which may have been moved in antiquity (or, indeed, in modern times) from its original location. This study goes hand in hand with the analysis of the typology of funerary monuments (and the reliefs that decorate them), the evolution of letterforms or the formulae used on epitaphs and on official decrees.[8] Second, subjecting the abovementioned formulae to an in-depth, linguistic study may point to chronological or local variations.[9] Finally, knowledge of dialectal variants is fundamental to the restitution of fragmentary or incomplete texts.[10]

Only systematically classifying the breadth and variety of linguistic variables can determine whether we are dealing with a real linguistic change, or whether they are cases of spelling mistakes, or other forms of graphic phenomena. Analysis of particular spelling variations associated with 'low' texts (including private letters, *defixiones*, etc.) vis-à-vis canonical spelling rules may, in addition, help us to determine linguistic phenomena across different sociolinguistic environments.

Linguistic markup of epigraphical corpora

Greek alphabetical inscriptions from Ancient Crete share the identifying characteristics of a corpus that can be used for linguistic annotation:[11]

- Textual files in Unicode (see below for some specific problems with Unicode fonts).
- Data from natural communicative situations, viz. data that were not intended to form part of a corpus.
- A certain balance, in the sense that the subgroups constitute a representative sample of the total corpus.

[3] The syllabic documents in Greek (Linear B) are not considered in this project as they exhibit specific characteristics and belong to a particular branch of Greek epigraphy.
[4] Ferrara 2010: 13-16.
[5] The question of the unity of the so-called 'Cretan dialect' has been addressed by Brixhe and Bile (1991).
[6] Cretan documents from the 4th century BC are rare.
[7] Brixhe 1993.
[8] Reinach 1885: 237-239.
[9] Robert 1961: 484.
[10] Guarducci 1967: 21.
[11] Gries and Berez 2017.

```
<sentence id='1' document_id='' subdoc='' span=''>
  <word id='1' form='Ἐπιφίλα' lemma='Ἐπιφίλα' postag='n-s---fn-' relation='SBJ' head='8' gloss='Epiphila' sg='nmn dpd'/>
  <word id='2' form='Σώσω' lemma='Σῶσος' postag='n-s---mg-' relation='ATR' head='1' sg='gnt dpd prp pss_bln' gloss='Sosos'/>
  <word id='3' form='Ἐπιθέτωι' lemma='Ἐπίθετος' postag='n-s---md-' relation='OBJ' head='8' sg='dtv dpd prp int adv' gloss='Epithetos'/>
  <word id='4' form='Θαρσαγόρα' lemma='Θαρσαγόρας' postag='n-s---mg-' relation='ATR' head='6' sg='gnt dpd prp pss_bln' gloss='Tharsagoras'/>
  <word id='5' form='ἀ' lemma='ὁ' postag='l-s---fn-' relation='ATR' head='6' sg='' gloss='the'/>
  <word id='6' form='γυνὰ' lemma='γυνή' postag='n-s---fn-' relation='APOS' head='1' sg='nmn dpd' gloss='wife'/>
  <word id='7' form='μναμεῖον' lemma='μνημεῖον' postag='n-s---na-' relation='OBJ' head='8' sg='acc dpd aff' gloss='memorial'/>
  <word id='8' insertion_id='0007e' artificial='elliptic' head='0' relation='PRED' form='ἐπέθηκε' gloss='dedicate' lemma='ἐπιτίθημι' postag='v3saia---'/>
</sentence>
```

Figure 1: Automatic *Arethusa* morphological analysis of *Epithetos*' epitaph (*IC* I viii 29).

The metadata associated with the digital edition of inscriptions afford useful information regarding texts' linguistic and dialectal analysis. The origin, date, typology, alphabet used, and the type of writing, etc., provide important clues that can help us to recognize particular trends regarding the geographical and/or chronological concentration of certain linguistic features.

Linguistic annotation of Ancient Greek: POS and lemmatisation

The practice of linguistically annotating textual corpora has developed over recent decades. Different kinds of analysis can be performed. These forms of analysis can comprise fundamental units (e.g. phonemes) or more complex units (e.g. phrases), to semantic and lexical annotation. Like any other form of semantic markup, the aim of linguistic annotation is to provide a textual corpus that might be useful for further research. Linguistic annotation adopts a semantic annotation language (for instance, XML), which invokes specific elements and attributes.

With regard to ancient languages, various methods of POS (parts of speech) encoding and lemmatisation, which contain linguistic information, are available: *Index Thomisticus Treebank* (IT-TB); *PROIEL Treebank*; *SEMATIA Treebank*; and *Ancient Greek and Latin Dependency Treebank* (AGLDT).[12] The different algorithms come in the form of a so-called Treebank, viz. a treelike diagram that signals information relating to parts of speech (POS) and the grammatical categories thereby associated, as well as the syntactical relations between the words of a text.

One of the most successful APIs that has been developed for the analysis of Greek and Latin is the *Ancient Greek and Latin Dependency Treebank* (AGLDT), integrated as an Arethusa application within the *Perseids* group of applications.[13] Arethusa moreover enables extensive and thorough syntactical annotation.[14] The most recent version (updated in 2018) enables the analysis of morphological and syntactical levels, as well as semantic functions.

The POS markup of a text from the dialectal Cretan corpus using Arethusa is straightforward. To illustrate, I have analysed (with the help of Arethusa) the epitaph of *Epithetos*:[15]

> Ἐπιφίλα Σώσω | Ἐπιθέτωι Θαρσαγόρα | ἀ γυνὰ μναμεῖον
> 'To *Epithetos*, son of *Tharsagoras*, her wife, *Epiphila*, daughter of *Sosos*, (sc. dedicated) a monument'.

Arethusa automatically recognizes most of the morphological features of the text. The annotated result is illustrated bellow in Figure 1.

We should emphasize that the markup schema of the XML of the AGLDT ('simple and intuitive', according to one of its creators)[16] does not follow the TEI guidelines, which is currently the most popular schema used in the digital edition of Greek and Latin epigraphical texts. This discrepancy entails that the linguistic markup must be kept in a different file, which can be validated with the particular schema. I will address this issue below.

Problems with conventional annotators

Morphological and syntactical analyses of the Cretan corpus with Arethusa faces two fundamental problems. The first is specific to the type of transcription (in itself, an encoding choice), which is used to transcribe the texts of the Cretan corpus. The second more generally concerns corpora with TEI-EpiDoc markup.

Cretan texts written in the local alphabet encounter font-based issues when they are digitally transcribed. For instance, the first three lines of the so-called 'Gortyn Code' (mid-5th century BC),[17] can be transcribed using decomposed or precomposed characters, i.e. 'a character that is equivalent to a sequence of one or more other characters'.[18] Figure 2 reproduces the first three lines of the 'Gortyn Code' with decomposed or precomposed characters using the font IFAO Unicode:

[12] Celano 2019b.
[13] The *Perseids* project (https://www.perseids.org).
[14] The *Ancient Greek and Latin Dependency Treebank* (AGLDT).
[15] *IC* I viii 29 (Cantanos, 2nd century BC). The attribution to Cnossus by M. Guarducci is unlikely.
[16] Celano 2019.
[17] *IC* IV 72.
[18] Decomposable Character (http://www.unicode.org/glossary/#decomposable_character).

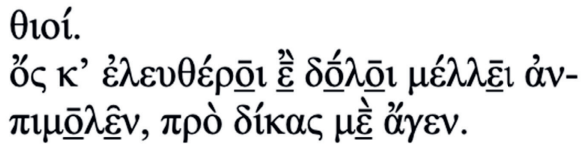
θιοί.
ὅς κ’ ἐλευθέρ̲ο̄ι ε̲̃ δ̲ό̲λ̲ο̄ι μέλλ̲ε̄ι ἀν-
πιμ̲ο̄λ̲ε̃ν, πρὸ δίκας μ̲ε̃ ἄγεν.

Figure 2: First three lines of the ‘Gortyn Code’ (*IC* IV 72) with decomposed or precomposed characters (IFAO Unicode font).

In the text of Figure 2, ἄ, ἀ, έ, ί, ὅ, ὸ, are Unicode decomposed characters that are normally used in the transcription of Greek texts and are recognized by Arethusa. However, the underlined letters are likewise decomposed characters, but which are currently only available in the so-called ‘Private Areas’ of some Unicode fonts (e.g. New Athena Unicode, IFAO Unicode, etc.). The latter are not recognized by Arethusa, and automatic POS analysis is therefore not possible. The analysis in Arethusa is possible only if the markup is performed manually and the lexical items are considered as sequences of identifiable and unidentifiable characters (e.g. ε̃, δ̲ό̲λ̲ο̄ι and μέλλ̲ε̄ι),[19] creating specific lemmata that do not correspond with the available dictionary entries (e.g. ἥ, δούλος and μέλλω).

On the other hand, the Cretan epichoric alphabet uses particular letters, e.g. *san* instead of *sigma* = σ, <S> instead of *iota* = ι. This is, in most cases, unproblematic because they are transcribed using conventional letters of the Greek alphabet (e.g. *sigma, iota*, etc.). However, in the Cretan local alphabet, unvoiced and aspirated labial and velar stops are written with the same letters, i.e., π represents φ and π, whereas κ represents χ and κ. In the same vein, 5th-century Cretan inscriptions, like the ‘Gortyn Code’, do not differentiate between long and short vowels. Finally, archaic (but not exclusively) inscriptions render nasals before labial stops with nu (ν), whereas in Classical Greek, mu (μ) is compulsory. Epigraphists typically maintain in their transcriptions of Cretan inscriptions all of these particularities. Consequently, in the transcribed form ἀνπιμ̲ο̄λ̲ε̃ν in the above example (Figure 2), we find the following equations: π = φ; ο̄ = ω; ε̃ = η; ν = μ.

Furthermore, digital epigraphical texts with TEI-EpiDoc markup (editorial interventions, line separations, etc.) cannot be directly encoded with AGLDT markup in Arethusa: only plain Unicode text is permitted.

[19] In order to illustrate my point, I leave unchanged the unidentifiable characters.

Analysis of linguistic variation of texts with editorial markup

Having taken into consideration the final problem mentioned in the previous section, several ongoing projects are in the process of developing methods to overcome the obstacles faced by texts already annotated with editorial markup. The field of papyrology has been especially dynamic in this area. In particular, researchers from the project *Sematia* (a Finnish research program) have been focusing on the linguistic encoding of documentary papyrus. *Sematia* has conceived of an automatic process to extract the text of an encoded EpiDoc edition (e.g. texts from the online database, *papyri.info*). The *Sematia* extraction process creates two different textual versions: the first is a sort of ‘diplomatic’ edition, in which words are separated. This produces the first textual version, which is denominated ‘original’ edition. The second (named ‘standard’) likewise contains the editorial interventions. A comparison between the two ‘editions’ produces a third layer that allows for the analysis of variants.[20]

Researchers from *Sematia* have also proposed the use of a new element <var>, as part of a specific XML schema for the study of papyrological linguistic variation.[21] Theoretically, this approach could be adopted for the analysis of texts of the e-SEC corpus. However, the creation of a particular schema goes against an established encoding consensus shared by the digital epigraphic community.

The database *Trismegistos* has also recently come up with the tool ‘Trismegistos Text Irregularities’, which according to its developers ‘is a tool for the collection and analysis of examples of linguistic variation.’[22] Based on the editorial markup of texts in the Papyrological Navigator of *papyri.info* database it begins with regularizations included in papyri by the editors using the elements <reg> and <orig>, children of <choice>. For instance, in the digital edition of a will dated to 284 BC (*P.Eleph.* 2), the infinitive τραφεῖν (line 12) has been encoded in the following way:[23]

```
<choice>

    <reg>τρέφειν</reg>
    <orig>τράφειν</orig>

</choice>
```

[20] Vierros and Henriksson 2017; Vierros 2018.
[21] Vierros and Henriksson 2017.
[22] Trismegistos Text Irregularities (https://www.trismegistos.org/textirregularities/methodology.php).
[23] *P.Eleph.* 2 (http://papyri.info/ddbdp/p.eleph;;2dupl).

The comparison between the 'original' forms and their 'regularization' creates an interesting subset of raw data, which contains information about linguistic variations in papyri.

As per EpiDoc Guidelines, '[i]f your project makes a distinction between text corrected as a result of scribal error [...] and text normalized or regularized from a dialect or phonetic spelling, grammatical form, etc., [...] then these normalizations can be tagged with <orig> (the original, scribal form) and <reg> (the regularized form)'.[24] In the TEI Guidelines, <reg> and <orig> are introduced as a way of regularizing non-standard spellings: 'When the source text makes extensive use of variant forms or non-standard spellings, it may be desirable for a number of reasons to regularize it: that is, to provide "standard" or "regularized" forms equivalent to the non-standard forms'. However, the TEI guidelines add that 'typical applications for these elements include the production of editions intended for student or lay readers, linguistic research in which spelling or usage variation *is not the main question at issue*, production of spelling dictionaries, etc.' (my emphasis).[25] As aforementioned, the main question at issue within the e-SEC project are variations regarding spelling and usage.

The use of the 'Trismegistos Text Irregularities' variation tool for the analysis of dialectal epigraphical Greek texts is problematic. First, it entails the adoption of a similar editorial practice for inscriptions. For instance, in the epitaph of Epithetos (see above), all dialectal forms should be marked-up with <reg> and <choice>. Although this may be a practice within the field of digital papyrology, digital editions of dialectal inscriptions have not adopted it.[26] In fact, <reg> and <orig> are not markers of analysis nor markers of an interpretation of a text; instead, they constitute a deliberate editorial strategy.

The 'Trismegistos Text Irregularities' tool moreover poses another problem. In the case of the first lines of the 'Gortyn Code', the dialectal active present infinitive ἀνπιμῶλῆν ('take legal action against someone concerning something') does not have an equivalent in standard Greek. This is because the verb is only used in Cretan texts. The 'regularization' of the form (as if ἀμφιμωλέω existed in the standard Greek) is scientifically inappropriate and wrong-headed.

In sum, then, while the 'Trismegistos Text Irregularities' may contribute to the analysis of already encoded documentary papyrus, it cannot be adopted as an analytical tool of linguistic variations within a digital corpus of Ancient Greek inscriptions. Indeed, as Stolk well notes, '[t]he traditional method of regularization is not suitable to encode variation consistently and objectively'.[27]

Stand-off markup

Another approach, which seeks to overcome the obstacles associated with the linguistic markup of corpora with EpiDoc editorial encoding, is discussed in the guidelines of TEI (chapter 16.9), as well as by Celano.[28] Two different methods of encoding are envisioned as follows:

- Internal or inline markup: viz. markup that is already present in an XML source document.
- Stand-off markup: the source text remains independent in a file and the annotation takes place elsewhere. The source document can either be an XML file or a simple text. Its encoding is linked to the source text through pointers, following the same principle that applies to images or sounds associated with a text.

This type of encoding betrays certain advantages. First, the starting point remains a 'blank edition', which does not contain editorial marks and is the closest to the original text. Furthermore, multiple layers of encoding can be added depending on the project's specific preoccupations. For instance, in the field of linguistics, phonology can be distinguished from morphology and syntax.

Annotation and dialectology

Traditionally, the linguistic analysis of dialectal variants of Ancient Greek has involved the description of a fairly broad number of linguistic phenomena in a given period of its diachronic evolution. The description of a specific linguistic variety constitutes an inventory of its phonetic, morphological, syntactical and lexical features. In this chapter, I will only address questions concerning the markup of phonological and morphological traits.

The phonological features of an historical corpus (of not merely dead languages) are always associated with the graphic system (sc. alphabet, syllabary, etc.), as well as the attested modifications of the same words through time and space. For instance, some Ancient Greek dialects exhibit -ρ instead of -ς for the final /s/

[24] EpiDoc. Regularization (http://www.stoa.org/epidoc/gl/latest/trans-regularization.html).
[25] Text Encoding Initiative. Regularization and Normalization (https://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#COEDREG).
[26] See the encoding of dialectal inscriptions of the digital Cyrenaean corpus IGCyr (https://igcyr.unibo.it/).
[27] Stolk 2018: 135.
[28] Celano 2019a.

sound. The graphic variants of a particular geographical area in terms of the standard norm, or other specific variants in other regions constitute the core of the description of the phonological features of a dialect. Thus, in the standard handbook on the language of Attic inscriptions, the phonological description includes 28 graphic variants for the system of simple vowels (both long and short).[29] Similarly, the standard phonology of Greek papyri of the Ptolemaic period has 27 sections.[30]

In the case of the dialectal variants of Crete, studies organize the material in different ways.[31] Some handbooks also have a synthetic presentation of the relevant linguistic features for dialectal classification.[32] One such presentation of dialectal features is particularly associated with epigraphic corpora: the *Index grammaticus*. In the first two editions of his collection of Greek inscriptions, W. Dittenberger included (for the first time, as far as I am aware) a list of linguistically important variants found in dialectal epigraphical documents.[33] More recent volumes of the collection of *Inscriptiones Graecae* also include this type of index. The *Index grammaticus* is not exclusive to dialectal corpora *stricto sensu* (i.e. with inscriptions written in one of the different Ancient Greek dialects), and may also appear in collections of late texts written in the standard Greek variant.[34] The rationale behind this type of index is simple: forms attested in inscriptions are categorized according to linguistic phenomena presented in a simple manner. For instance, 'ᾱο > ᾱ: Ἀγησίλας, Λαμέδων, Λαχάρης' indicates that the vocalic contraction of /a:/ + /o/ is illustrated in the corpus by the three mentioned forms.

Let us now examine what variants can be analysed in a dialectal inscription. Again, I illustrate the example with *Epithetos*' epitaph, in which the following features are relevant to the dialectal analysis (the emphasized letters representing the graphic variation):

- /a:/ (Ἐπιφίλα, ά, γυνά, μναμεῖον) vis-à-vis /ɛ:/ in Attic-Ionic
- ο + ο) /ɔ:/ (Σώσω) vis-à-vis /o:/ in Attic-Ionic
- a + ο) /ɔ:/ (Σώσω) vis-à-vis /ao/ in other areas
- /ɔ:i̯/ word-finally (Ἐπιθέτωι) vis-à-vis /ɔ:/ in the same area
- ᾱ + ο) /a:/ (Θαρσαγόρα) vis-à-vis the non-regular /o:/ in Attic-Ionic and /a:o/, /a:u/ in other areas

Let us now turn to morphology. The POS markup, to this end, could theoretically be used for the purposes of annotating Ancient Greek dialectal features, since the traditional POS analysis contains the linguistic information that dialectal analysis uses regularly in the classification of morphological features. However, in the dialectal classification of these features, different markers of morphological categories are essential to the task of classification. Unfortunately, POS markup does not indicate whether a noun or a name belongs to a particular flexional type.

We can illustrate the above with two examples taken from the first lines of the 'Gortyn Code' (Figure 2). The lexical unit ἀνπιμῶλῆν is an active infinitive of a contracted verb: the ending -ῆν likely represents the local variant (-/ɛ:n/) vis-à-vis the standard form -εῖν (-/e:n/) in Attic-Ionic. On the other hand, the lexical unit ἄγεν is also an active infinitive, albeit of a non-contracted verb. In this case, the ending -εν represents the local variant (-/en/) vis-à-vis the standard form -εῖν (-/e:n/) in Attic-Ionic. These two different verbal terminations are fundamental to the description of dialectal variants in Ancient Greek.

From a practical point of view, an XPath query of the content of the @postag attribute of the AGLTD schema could help us to classify all of the infinitives. However, leaving aside the fact that this is merely a shortcut (and *not* a dialectal approach to markup), the XPath query will result in a list that also includes the athematic present infinitives, which have specific dialectal characteristics (mainly, the final -μεν vis-à-vis Attic-Ionic -ναι).

The element <distinct>

The TEI schema includes the element <distinct>, which according to its Guidelines, 'identifies any word or phrase which is regarded as linguistically distinct, for example as archaic, technical, *dialectal*, non-preferred, etc., or as forming part of a sublanguage' (my emphasis).[35] The TEI Guidelines further recommend that 'lemmatized words or expressions can be encoded with specific characteristics from a chronological, geographical or social point of view' (3.3.2.3 Other Linguistically Distinct Material).

Four attributes are specific to <distinct>: @time; @social; @type; and @space. The lattermost attribute, @space, 'specifies how the phrase is distinct diatopically', whereas @type 'specifies the sublanguage or register to which the word or phrase is being assigned'. On my view, this element could prove useful in the case of texts incorporating, for instance, a range

[29] Threatte 1980: xv-xvi.
[30] Mayser and Schmoll 1970: vi-viii.
[31] Bechtel (1923) has 52 sections; Thumb and Kieckers (1932), 37; and Buck (1955), 38.
[32] Buck's (1955) famous 'Chart I' is included at the end of his handbook.
[33] Dittenberger 1883: 780-785; Dittenberger 1898-1901: vol. III, 224-240. The editors of the posthumous 3rd edition (Dittenberger 1915-1924) regrettably decided to dispense with this section of the indices.
[34] See Hallof and Matthaiou 2003; Summa 2011; Nigdelis 2017. The *Index grammaticus* in these corpora was compiled by J. Curbera.
[35] TEI Guidelines Version 4.0.0 (https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-distinct.html) .

```
<encodingDesc>
  <editorialDecl>
      <interpretation>
          <p>Ce fichier contient une analyse dialectale du Code de Gortyne

              <interpGrp type="dialecte_classement">
                  <interp xml:id="ST">Forme non dialectale</interp>
                  <interp xml:id="D">Forme dialectale</interp>
              </interpGrp>

              <interpGrp type="dialecte_phonologie">
                  <interp xml:id="E-I-1">iota pour epsilon devant
                      voyelle</interp>
                  <interp xml:id="E-I-0">epsilon devant voyelle
                      inchangé</interp>
                  <interp xml:id="EEE">contraction e + e long</interp>
                  <interp xml:id="h0">psilose</interp>
                  <interp xml:id="h1">aspiration initiale</interp>
                  <interp xml:id="V0-A">élision A</interp>
                  <interp xml:id="OOIf-1">diphthongue finale ωι</interp>
                  <interp xml:id="OOIf-0">diphthogue finale ω</interp>
                  <interp xml:id="EEIf">diphthongue finale ηι</interp>
                  <interp xml:id="A_H-1">alpha ancien conservé</interp>
                  <interp xml:id="A_H-0">alpha ancien non conservé</interp>
              </interpGrp>

              <interpGrp type="dialecte_morphologie">
                  <interp xml:id="INF_ACT_EW">infinitif actif verbe -έω</interp>
                  <interp xml:id="INF_ACT_W">infinitif actif verbe -ω</interp>
              </interpGrp>

               </p>
      </interpretation>
  </editorialDecl>
</encodingDesc>
```

Figure 3: List of features in the interpretation declaration.

of variants from the standard language (Koine) and the Cretan dialect.

Linguistic categories in TEI

As aforementioned, the objective of the dialectal markup is to create a corpus containing grammatical information relevant to the description of the linguistic variation of a specific language. As I have shown, in the case of epigraphical corpora, the *Index grammaticus* offers an elegant 'codification' of this linguistic information, as arranged in lists which categorizing lexical items according to shared characteristics. It strikes me that we can adapt the 'code' of the epigraphical *Index grammaticus* by means of a specific group of elements discussed in the chapter 'Lightweight Linguistic Annotation' (17.4.2) of the TEI Guidelines.[36]

The basic linguistic element of the TEI schema is <w> (see 17.1 'Linguistic Segment Categories' and 17.4 'Linguistic Annotation'), which is already used for lemmatisation purposes by projects heeding the EpiDoc Guidelines. Stolk, for instance, has proposed to combine the use of three attributes of the att.linguistic category of the

```
<p>
    <w ana="#D #E-I">θιοί</w>.
    <w ana="#h0">ὅς</w>
    <w ana="#V0">κ’</w>
    <w ana="#OOI">ἐλευθέρῳι</w>
    <w>ἐ</w>
    <w ana="#OOI">δόλῳι</w>
    <w ana="#EEI">μέλλῃι</w>
    <w ana="#INF_ACT_EW #EEE">ἀνπιμῶλῆν</w>,
    <w>πρὸ</w>
    <w ana="#A_H">δίκας</w>
    <w>μὲ</w>
    <w ana="#INF_ACT_W">ἄγεν</w>
</p>
```

Figure 4: Proposal of analysis of variants of the first lines of the 'Gortyn Code' (*IC* IV 72).

element <w>: @lemma; @pos (part of speech); and @msd (morphosyntactic description).[37] In the case of Cretan dialectal inscriptions, Stolk's approach offers the possibility of creating a precise index, similar to the one created by Fraenkel (1915). However, applying this encoding strategy in a dialectal corpus may grow problematic, since the analysis is not rooted in morphological markers, but in a POS grammatical description.

[36] Text Encoding Initiative. Lightweight Linguistic Annotation (https://tei-c.org/release/doc/tei-p5-doc/en/html/AI.html#AILALW). See also Bański, Haaf and Mueller 2018.

[37] Stolk 2018: 135-136.

```
<p>
    <w lemma="θεός" ana="#D #E-I">θιοί</w>.
    <w lemma="ὅς" ana="#h0">ὅς</w>
    <w lemma="κα" ana="#V0">κ’</w>
    <w lemma="ἐλεύθερος" ana="#OOI">ἐλευθέρōι</w>
    <w lemma="ἤ">ἒ</w>
    <w lemma="δοῦλος" ana="#OOI">δόλōι</w>
    <w lemma="μέλλω" ana="#EEI">μέλλẽι</w>
    <w lemma="ἀμφιμωλέω" ana="#INF_ACT_EW #EEE">ἀνπιμōλẽν</w>,
    <w lemma="πρό">πρὸ</w>
    <w lemma="δίκη" ana="#A_H">δίκας</w>
    <w lemma="μή">μἒ</w>
    <w lemma="ἄγω" ana="#INF_ACT_W">ἄγεν</w>
</p>
```

Figure 5: Proposal of analysis of variants: internal encoding.

The TEI Guidelines have, in various ways, pre-empted this scenario, namely with a linguistic annotation that uses generic TEI devices elements (17.4.1). In addition to the specific attributes of att.linguistic family, the @ana attribute (analysis) indicates ‘one or more elements containing interpretations of the element on which the @ana attribute appears.’ This attribute, @ana, thereby allows for a particular linguistic interpretation of the element <w>.

In the manner of an *Index grammaticus*, whereby a definition is linked to different forms attested in the corpus (see above), the markup of a dialectal Greek inscription requires a list of features that can be linked to the attribute @ana. The specification of these features must be declared in the TEI Header under the element <editorialDecl>, which ‘is used to provide details of the editorial practices applied during the encoding of a text’ (2.3.3 The Editorial Practices Declaration). One of the elements embedded into the editorial declaration is <interpretation>, which provides a description of ‘the scope of any analytic or interpretive information added to the text in addition to the transcription’. The interpretations of different lexical items can be described under interpretation groups (element <interpGrp>), which ‘collect together a set of related interpretations which share responsibility or type’. Each feature can be specified in an element <interp>, which ‘summarizes a specific interpretative annotation which can be linked to a span of text’. Each lexical item with the attribute @ana can be linked to each <interp> via a conventional attribute @xml:id.

Illustrated in Figure 3 are a list of features specified in the interpretation declaration. In Figure 4, I further exemplify a possible analysis of variants of the first lines of the ‘Gortyn Code’:

Once the list of features has been established, an encoding practice must be chosen. Theoretically, two approaches are possible:

- Internal encoding. Every word is annotated with the element <w>, a common practice for the lemmatisation of words in a corpus. The attribute @ana can include as many interpretative notes as necessary, each one pointing to its definition or specification in the TEI Header (<interp>) (Figure 5).
- Instead of creating a new layer in the edition, it is possible to adopt another strategy: external markup. In this case, each lexical item, annotated with the element <w>, will have a unique identifier (the attribute @xml:id). In a separate file, each word is analysed, adding the interpretative notes (@ana) and the link to it (@target) (Figure 6).

Conclusions

By way of conclusion, I believe that the perspective of the e-SEC project can contribute to the current efforts of linguistic encoding Ancient Greek texts. This short contribution has shown that the TEI offers a robust system for developing specific encoding practices to capture dialectal variation. I have proposed to explore two complementary strategies. The first consists in the incorporation of the element <distinct> as a way of annotation of linguistic variants. Second, the attribute @ana of the element <w>, in connection with an editorial declaration (<editorialDecl>), in which a detailed list of features is specified with the element <interpretation>, might offer a practical solution regarding the markup of dialectal variants in epigraphical corpora.

Bibliography

Bański, P., Haaf, S. and Mueller, M. 2018. Lightweight grammatical annotation in the TEI: new perspectives, in N. Calzolari, Kh. Choukri, Chr. Cieri, Th. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga (eds) *Proceedings of the*

```
<p>
    <lb/>
    <w xml:id="word-1">θιοί</w>.
    <lb/><w xml:id="word-2">ὅς</w>
    <w xml:id="word-3">κ</w>’
    <w xml:id="word-4">ἐλευθέρōι</w>
    <w xml:id="word-5">ἐ</w>
    <w xml:id="word-6">δόλōι</w>
    <w xml:id="word-7">μέλλῆι</w>
    <w xml:id="word-8">ἀν<lb break="no"/>πιμōλῆν</w>,
    <w xml:id="word-9">πρὸ</w>
    <w xml:id="word-10">δίκας</w>
    <w xml:id="word-11">μὲ</w>
    <w xml:id="word-12">ἄγεν</w>.
</p>
```

```
<linkGrp type="dialectal-annotation">
    <link target="#word-1 #D #E-I"/>
    <link target="#word-2 #h0"/>
    <link target="#word-3 #V0-A"/>
    <link target="#word-4 #OOIf"/>
    <link target="#word-6 #OOIf"/>
    <link target="#word-7 #EEIf"/>
    <link target="#word-8 #INF_ACT_EW #EEE"/>
    <link target="#word-10 #A_H"/>
    <link target="#word-12 #INF_ACT_W"/>
</linkGrp>
```

Figure 6: Proposal of analysis of variants: external encoding.

eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018. Miyazaki, Japan: 1795-1802. Paris: European Language Resources Association (ELRA).

Bechtel, F. 1923. *Die griechischen dialekte 2: Die westgriechischen Dialekte*. Berlin: Weidmann.

Brixhe, Cl. 1993. Le declin du dialecte crétois: essai de phénoménologie, in E. Crespo, J.L. García-Ramón and A. Striano (eds) *Dialectologica Graeca. Actas del II coloquio internacional de dialectologia griega (Miraflores de la Sierra, 19-21 de junio de 1991)*: 37-71. Madrid: Ediciones de la Universidad Autónoma de Madrid.

Brixhe, Cl. and Bile, M. 1991. Le dialecte crétois. Unité ou diversité ?, in Cl. Brixhe (ed.) *Sur la Crète antique. Histoire, écritures, langues*: 85-127. Travaux et mémoires. Études anciennes 6. Nancy: Presses Universitaires de Nancy.

Buck, C.D. 1955. *The Greek Dialects. Grammar, Selected Inscriptions, Glossary*. Chicago: University of Chicago Press.

Celano, G.G.A. 2018. An automatic morphological annotation and lemmatization for the IDP papyri, in N. Reggiani (ed.) *Digital Papyrology II: Case Studies on the Digital Edition of Ancient Greek Papyri*: 139-148. Berlin: De Gruyter.

Celano, G.G.A. 2019a. Standoff annotation for the Ancient Greek and Latin dependency treebank, in *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019), May 8-10, 2019, Brussels*: 149-153. New York: Association for Computing Machinery.

Celano, G.G.A. 2019b. The dependency treebanks for Ancient Greek and Latin, in M. Berti (ed.) *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*: 279-297. Grundfragen der Informationsgesellschaft 10. Berlin – Boston: De Gruyter.

Depauw, M. and Stolk, J.V. 2015. Linguistic variation in Greek papyri: towards a new tool for quantitative study. *Greek, Roman and Byzantine Studies* 55: 196-220.

Dittenberger, W. 1883. *Sylloge inscriptionum Graecarum*. Leipzig: Hirzel.

Dittenberger, W. 1898-1901. *Sylloge inscriptionum Graecarum* (2nd edition). Leipzig: Hirzel.

Dittenberger, W. 1915-1924. *Sylloge inscriptionum Graecarum* (3rd edition). Leipzig: Hirzel.

Ferrara, S. 2010. Mycenaean Texts: The Linear B Tablets, in E.J. Bakker (ed.) *A Companion to the Ancient Greek Language. Blackwell Companions to the Ancient World. Literature and Culture*: 11-24. Chichester/Malden, MA: Wiley-Blackwell.

Fraenkel, E. 1915. Index der kretischen Inschriften, in H. Collitz and O. Hoffmann (eds) *Sammlung der griechischen Dialekt-Inschriften 4.4*: 1029-1208. Göttingen: Vandenhoeck and Ruprecht.

Gries, S.Th. and Berez, A.L. 2017. Linguistic annotation in/for corpus linguistics, in N. Ide and J. Pustejovsky (eds) *Handbook of Linguistic Annotation*, vol. I: 379-381. Dordrecht: Springer.

Guarducci, M. 1935-1950. *Inscriptiones Creticae I-IV*. Rome: Libreria dello Stato.

Guarducci, M. 1967. *Epigrafia greca I*. Rome: Istituto poligrafico dello Stato.

Hallof, K. and Matthaiou, A. 2003. *IG XII 6. Inscriptiones Chii et Sami cum Corassiis Icariaque*. Berlin: De Gruyter.

Keersmaekers, A. 2020. Creating a richly annotated corpus of papyrological Greek: the possibilities of natural language processing approaches to a highly inflected historical language. *Digital Scholarship in the Humanities* 35(1): 67-82.

Mayser, E. and Schmoll, H. 1970. *Grammatik der griechischen Papyri aus der Ptolemäerzeit* 1: *Einleitung und Lautlehre*. Berlin: De Gruyter.

Nigdelis, P.M. 2017. *IG X 2 1 Inscriptiones Thessalonicae et viciniae. Supplementum*. Berlin: De Gruyter.

Reinach, S. 1885. *Traité d'épigraphie grecque*. Paris: Ernest Leroux.

Robert, L. 1961. Les épigraphies et l'épigraphie grecque et romaine, in *L'Histoire et ses méthodes. Encyclopédie de la Pléiade* 11: 453-497. Paris: Gallimard. (= Robert,

L. (2007). *Choix d'écrits*: 87-114. Paris: Les Belles Lettres).

Stolk, J.V. 2018. Encoding linguistic variation in Greek documentary papyri, in N. Reggiani (ed.) *Digital Papyrology II: Case Studies on the Digital Edition of Ancient Greek Papyri*: 119-138. Berlin: De Gruyter.

Summa, D. 2011. *IG IX 1² 5. Inscriptiones Locridis orientalis*. Belin: De Gruyter.

Threatte, L. 1980. *The Grammar of Attic Inscriptions* 1: *Phonology*. Berlin – New York: De Gruyter.

Thumb, A. and Kieckers, E. 1932. *Handbuch der griechischen Dialekte*. Heidelberg: Winter.

Vierros, M. 2018. Linguistic annotation of the digital papyrological corpus: Sematia, in N. Reggiani (ed.) *Digital Papyrology* II: *Case Studies on the Digital Edition of Ancient Greek Papyri*: 105-118. Berlin: De Gruyter.

Vierros, M. and Henriksson, E 2017. Preprocessing Greek papyri for linguistic annotation, *Journal of Data Mining and Digital Humanities. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages* (https://jdmdh.episciences.org/paper/view?id=1385).