# Epigraphic Treebanks: Some Considerations from a Work in Progress

**Francesca Dell'Oro and Giuseppe G. A. Celano**

**Pre-Published for FirstDrafts@Classics@, July 26, 2019***

## Abstract

The guidelines of the *Ancient Greek Dependency Treebank 2.0* have been written to annotate Ancient Greek texts. The epigraphic texts, however, pose a challenge for those carrying out morphosyntactic annotation: should we remain as close as possible to the actual epigraphic text, or represent it in an interpreted and normalized version? How should all epigraphic peculiarities which do not have standard editorial representation, such as, for example, punctuation marks, be treated? A small corpus such as that of the inscriptions of the Euboean colonies of Sicily of the archaic and classical period has allowed us to test different options and evaluate the annotation challenges. This contribution is the result of a discussion about the advantages and disadvantages of often opposed annotation possibilities.

We present here our first proposal for an adaptation of the guidelines to analyse the morphosyntax of inscriptions, which we hope will stimulate discussion between epigraphists and linguists. In particular, we propose to try to stick to the epigraphic evidence as far as possible and therefore render its complexities (e.g., local alphabets, dialect variants not attested in literary texts, ellipsis, punctuation marks, and word forms which can be linguistically interpreted differently), while trying to preserve consistency with the annotation of literary texts.

* This paper was accepted for publication in the proceedings of the DHANT conference (= Digital Humanities and Antiquity, Grenoble, 2-4 September 2015) at the end of 2016.

## 1. Goal and background

The main goal of this paper is to stimulate discussion about linguistic annotation of epigraphic texts by pointing out some of the problems we encountered while annotating Greek inscriptions according to the guidelines drawn up by Giuseppe G. A. Celano (2014) to parse Greek texts (*Ancient Greek Dependency Treebank 2.0*). These guidelines expand the *Guidelines for the Syntactic Annotation of the Ancient Greek Dependency Treebank (1.0)*, by refining morphosyntactic annotation and adding a third, semantic, layer[1]. Our aim for the near future is to create a complement to the already existing *Guidelines* in order to support the specificities of annotating inscriptions.

In a recent paper, Francesca Dell'Oro (2015) showed that epigraphic texts often constitute neglected corpora in the study of synchronic and diachronic syntax. This could be due to some inherent characteristics (e.g., their usual shortness in comparison with literary texts, the usual impossibility of identifying the author, the repetitive presence of formulaic language, and so on) as well as to their scarce accessibility. Starting from such considerations about inscriptions, we are trying to adapt the *Guidelines*, by identifying the peculiarities of epigraphical texts which are significant for linguistic annotation. We have also tried to consider the relevant differences that specialists of different disciplines (epigraphy, papyrology, palaeography, etc.) show in editorial practices and in their approaches to ancient texts. As testing corpus, we chose the Euboean inscriptions[2]. More specifically, the texts presented in this paper come for the most part from the collections of Greek dialectal inscriptions from Sicily edited by Laurent Dubois (*IGDS* I and *IGDS* II). These texts are short and not very complex from a syntactic point of view, but they present the

[1] The *Guidelines* improve those of the previous version (*AGDT 1.0*), in particular by adding a third layer of analysis, the advanced syntax/semantic layer, to the morphological and the (Prague) syntactic ones. Therefore, the *AGDT 2.0* is organized into three layers: the morphological layer, the (Prague) syntactic layer, and the advanced syntax/semantic layer based upon Smyth, 1920.

[2] Data will be made available for queries through the search application Tündra (http://weblicht.sfs.uni-tuebingen.de/Tundra/).

typical problems of epigraphic texts. The tool we used for annotation is Arethusa[3], which is already used for annotating a variety of texts transmitted through manuscripts and papyri and for the Marmor Parium inscription[4].

In the following paragraphs, we will present some of the problems we encountered and some of the possible solutions we thought of. In this way, we hope to encourage discussion about these issues in order to help annotation of inscriptions and provide some suggestions about how to adapt the annotation guidelines to the case of inscription annotation.

## 2. Local alphabets and dialectal forms

Unlike the Marmor Parium, which is a Hellenistic inscription and in which the common alphabet is used[5], the dialectal inscriptions of *IGDS* are often written in a local alphabet[6]. The Euboean inscriptions of this corpus are usually written without specific letter signs for the sounds which are later rendered through eta (<H>), omega (<Ω>) and the digraphs epsilon + iota (<EI>) and omicron + iota (<OY>). For example, the sign <E> can thus be used for rendering the short mid front vowel (which would correspond to <E> in the common alphabet), as well as the long open mid front vowel and the long close mid front vowel (which would correspond respectively to <H> and <EI> in the common alphabet).

---

[3] Arethusa (http://www.perseids.org/tools/arethusa/app/#/) is an open annotation environment, which is specifically designed for the annotation of Ancient Greek and Latin morphosyntax.

[4] See http://www.dh.uni-leipzig.de/wo/dmp/

[5] The traditional date for the adoption of the common alphabet is 403/402 BCE (Colvin 2007, p. 19). For clarity's sake and according to linguistic use, in this paper angle brackets are used when a letter of the alphabet (i.e. the grapheme) is being discussed. Slashes are used to denote a sound in broad phonetic transcription (i.e. the phoneme).

[6] As in the modern editions of literary texts, in inscriptions written in the common alphabet there are specific graphemes or combination of graphemes for the sounds /ɛ:/ (i.e. the long open mid front vowel which is rendered as <H>), /ɔ:/ (i.e. the long close mid back vowel which is rendered as <Ω>), /e:/ (i.e. the long close mid front vowel which is rendered as <EI>) and /o:/ (i.e. the long close mid back vowel which is rendered as <OY>). One should bear in mind that this praxis was not usual for inscriptions in the archaic and classical periods, when different local alphabets were in use.

We decided to transcribe the text without using the common alphabet and therefore attempting to interpret the actual word forms (as far as their spelling is concerned). This has the disadvantage that one cannot use some of the already existing resources to speed up the annotation process. In particular, we cannot use the Mate tagger (Celano 2016) and the morphological analyzer Morpheus, which mostly recognizes Ionic-Attic word forms. For example, if we introduce the first word of *IGDS* I 8 in its normalized form, i.e., Ζηνός (the genitive of Ζεύς), Morpheus will recognize the form automatically and will suggest the following: noun, singular, masculine, and genitive of the lemma "Ζεύς". On the other hand, if we introduce the word as it is written on the vessel (i.e., Ζενὸς), we would need to add such an annotation manually.

From a more general perspective, this annotation choice is directly linked to the question to what degree the annotator should work with the *epigraphic text* (i.e., the text as it is written on the object) rather than with the *edited text* (i.e., the interpreted epigraphic text). Once the sequence of letters ΖΕΝΟΣ (i.e., the epigraphic text) has been interpreted as morphologically equivalent to Ζηνός, it would be possible to transcribe and to annotate this last form. As it will be clear also from the other cases presented below, it is at this point that a fruitful discussion between different approaches can begin. Even though we do not want to suggest a definitive answer, it seems to us preferable that the annotator sticks to the epigraphic text: indeed, normalization of the spelling can often turn out to be very challenging, in that one original word form could potentially correspond to two or more normalized forms. This problem is usually triggered by the fragmentary nature of inscriptions where the linguistic context is missing.

Moreover, we decided to link the annotation of the inscriptions to the existing corpus via the lemma form and its morphological analysis. Once a morphological interpretation is provided, each word form is linked to an Ionic-Attic lemma[7]. For example, manuscripts

[7] It has to be added that annotating more than one morphological interpretation was out of scope, in that

attest the form Ζάγκλη for the ancient name of Messene, but inscriptions and coin legends attest the same name in the form Δάνκλη (e.g., *IGASMG* III 39). Morpheus cannot recognize Δάνκλη as the same lemma as Ζάγκλη. In this case, it is not necessary to introduce a second lemma: the annotator can specify that the lemma is the same, in that Ζάγκλη is taken to be a dialectal variant of Δάνκλη. This is in accordance with standard practice in non-dialectal Ancient Greek dictionaries (e.g. Liddell and Scott, 1996).

## 3. Implied words

A second important point we would like to focus on is that of implied words, especially when the main verb is missing in an inscription. Annotation of ellipsis is notoriously difficult to deal with, and there is no agreed strategy yet to annotate this phenomenon. On the one hand, one would like to avoid introducing elliptical nodes in a sentence, in that it is hard to provide rules which can consistently be applied by all annotators; on the other, it may be impossible to annotate a sentence according to our current annotation scheme without adding an elliptical node: this often holds true when the main verb of a sentence is missing.

For example, we introduce elliptical nodes, for inscriptions relating to ownership. Compare *IGDS* I 14a: [Η]εκαταίō ἓ[μί] "I am of Hekataios" (fig. 1) with *IGDS* I 14b : Ἡρακλείδα « Of Herakleidas » (fig. 2) in which the main verb is missing:

---

there is not yet agreement as to how to do this in an efficient way. We adopt the Perseus XML schema, which allows an elegant yet simple kind of morphosyntactic annotation. As for word forms, we had to decide whether to give priority to the original text or to the normalized one. However, one aim of a recently (2018) approved DFG-Project (http://gepris.dfg.de/gepris/projekt/408121292) is to provide stand-off annotation for the Ancient Greek and Latin Dependency Treebank. This will provide a solution for the problem of multiple annotations for the same token.


Fig. 1 *IGDS* I 14a: [H]εκαταίō ἓ[μί]

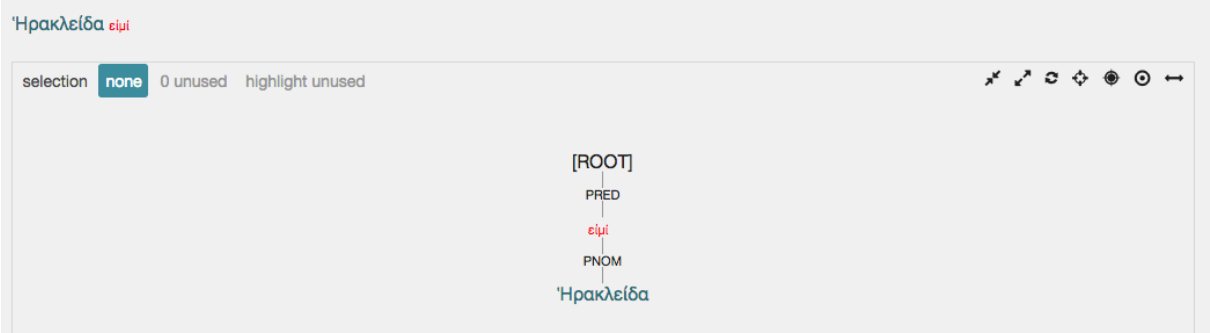


Fig. 2 *IGDS* I 14b: Ἡρακλείδα

The speaking-object formula led us to interpret *IGDS* 14b as « (I am) of Herakleidas », but - at least theoretically - we are not able to decide whether the writer meant this rather than « (This is) of Herakleidas ».

Another example from the corpus sheds further light on this point. The following inscription accompanies the dedication of parts of an armoury after a military victory: (IGDS I 4a) Διὶ [Ὀλ]υνπ̣ὶο̣(ι) Μεσσένιοι Λοκ[ρôν]. At first sight the translation "The Messenians (dedicated this leg-armour) of the Locrians to Olympian Zeus" could seem a fitting one, but, if we look at the formulation of other similar inscriptions, in particular IG I3 1467 (Olympia): Διὶ Ἀθēναῖοι Μέδōν **λαβόντες** (our emphasis) and IGASMG V 13a: σκῦλα **ἀπὸ Θουρίῶν** Ταραντῖνοι ἀνέθēκαν Διὶ Ὀλυμπίōι δεκάταν (our emphasis) « The Tarantines dedicated the spoils (taken) from the Thurians to Olympian Zeus as a tithe », we see that the

genitive Μέδων could be interpreted as governed by a missing participle λαβόντες through a missing preposition ἀπό: "After having taken (this leg-armour from) the Persians, the Messenians (dedicated it) to Olympian Zeus". Fig. 3 shows this interpretation. We have added an elliptical node not only for the main verb which could have been the aorist form ἀνέθηκαν, but also for the participle, which could have been λαβόντες (or another semantically similar verb), and the preposition ἀπό[8].

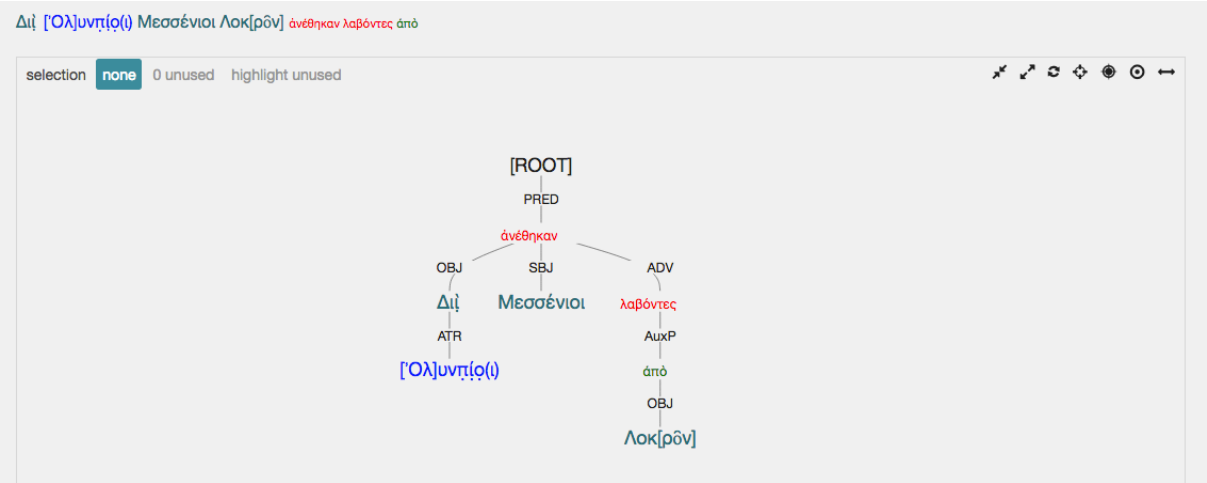


Fig. 3 *IGDS* I 4a

The elliptical nodes are always identifiable as something artificial, i.e, introduced by the annotator, both in the underlying XML structure and in the interface.

## 4. Word and phrasal division

Ancient Greek inscriptions were usually written without word separation (*scriptio continua*) and punctuation. Nevertheless, it is not unusual to find words and small units like phrases (rarely sentences) separated by punctuation marks which could take different forms (cf., e.g, the three points in *IGDS* I 1 discussed below)[9]. We suggest respecting the punctuation

[8] It would be useful to establish a common thesaurus for elliptical nodes.

[9] For punctuation in Greek inscriptions in general, cf. Guarducci, 1995, p. 391-397, and Jeffery, 1989, p. 50. For more information about the practice of punctuation in the various regions of Archaic Greece, cf. the section "Notes on letter-forms" for each local alphabet in the cited work by Jeffery.

originally available in the document as much as possible. We are aware of the problem that Unicode does not provide codepoints for each ancient punctuation mark, and so a desideratum for future research is to come up with a punctuation mark inventory and its Unicode rendering.

It is not yet completely clear what the function of punctuation was, but much of it seems to mark prosodic units (Morpurgo Davis 1987; Wachter 1999). This can be inferred, for example, when an enclitic or a proclitic word is present. Prosodic units can sometimes, but not always, coincide with syntactic units. In the corpus of Euboean inscriptions of Sicily, it is possible to find at least two examples of punctuation. The first one shows that ancient punctuation did not follow the same logic as modern punctuation.

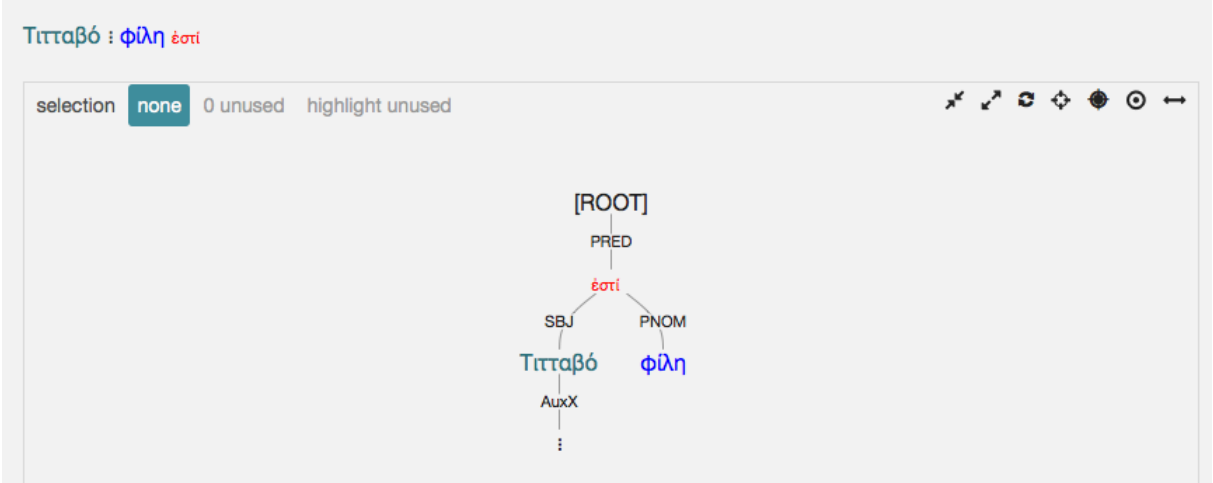


Fig. 4 *IGDS* I 1: Τιτταβό̋ ⁝ φίλη

In this case, we have interpreted the three points as a kind of comma ("Tittabó, she is dear (to me)") and we used the label AuxX, which is used for non-coordinating punctuation marks[10].

[10] The proper name Τιτταβώ could also be interpreted as an exclamation: "Tittabó! She is dear (to me)", but we have annotated here only the former interpretation.

Another interesting example is offered by *IGDS* I 11, which presents two sentences divided by two points: *IGDS* I 11: Εὐōπίδας *h*ιάλē Διεύχēς : / λοχαγὸς Δαῖτις[11]. The interpretation of this inscription is very uncertain. We can provide the following non literal translation following the interpretation suggested by Dubois in his commentary on this inscription : « Euopidas was sent, Dieuches (was sent). The commander (is) Daitis »[12].

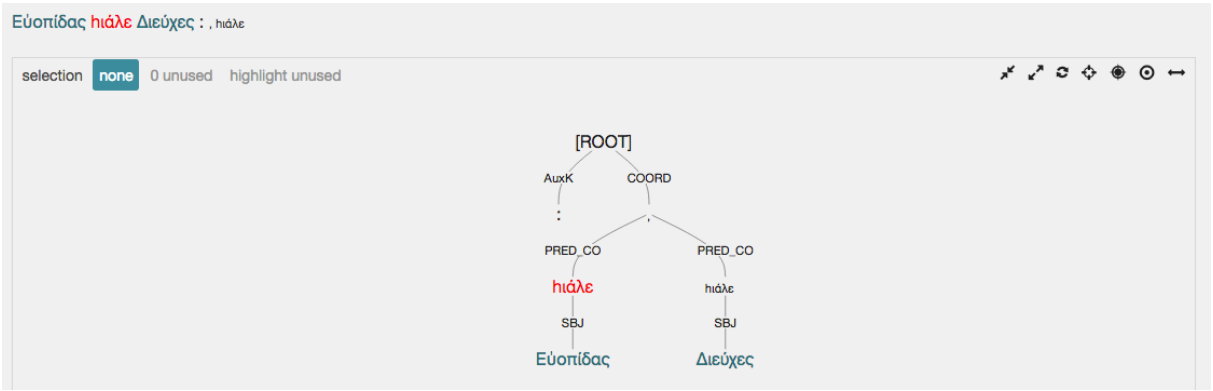


Fig. 5 *IGDS* I 11, 1: Εὐōπίδας ηιάλē Διεύχēς :

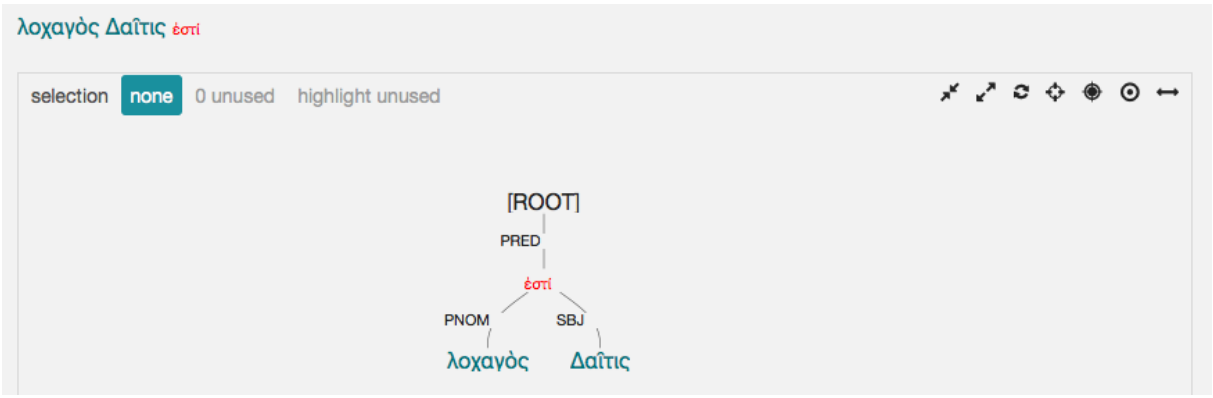


Fig. 6 *IGDS* I 11, 2: λοχαγὸς Δαῖτις

In this case, we have interpreted the punctuation mark as a kind of colon and attached it to the root of the tree (fig. 5), in accordance with the instructions in the *Guidelines* (cf. 2.11). We also added a comma in order to be able to annotate the coordination between the verbs.

[11] λοχαγὸς Dell'Oro : Λ/λοχαγος Dubois.

[12] For another suggestion, cf. *IGASMG* III 51.

## 5. Provisional conclusions

In this paper, we have presented some solutions which we have adopted to solve problems we encountered while annotating inscriptions. Because of their complexity, the annotation of these documents often requires ad-hoc instructions, which have not yet been offered by the available guidelines for (Classical) Ancient Greek. We have suggested transcribing the epigraphic text in the epigraphic alphabet (Section 2); reporting dialectal variants to a Ionic-Attic lemma (Section 2), which, together with morphological annotation, allows comparison of these texts with the others contained in the Ancient Greek Dependency Treebank; introducing elliptical nodes only when strictly needed (most notably, when the main verb is missing; Section 3); transcribing punctuation marks and trying to interpret them according to the *Guidelines* as much as possible (Section 4).

## Bibliography


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

CELANO Giuseppe G.A., Gregory Crane, and Saeed Majidi, 2016, « Part of Speech Tagging for Ancient Greek », *Open Linguistics* 2: 393-399.

COLVIN Stephen, 2007, *A historical Greek reader. Mycenaean to the Koiné*, Oxford, Oxford University Press, 2007.

DELL'ORO Francesca, « What Role for Inscriptions in the Study of Syntax and Syntactic Change in the Old Indo-European Languages? The *Pros* and *Cons* of an Integration of Epigraphic Corpora », dans C. Viti (dir.), *Perspectives on Historical Syntax*, Amsterdam, Benjamins, p. 271-290.

GUARDUCCI Margherita, 1995, *Epigrafia greca I. Caratteri e storia della disciplina. La scrittura greca dalle origini all'età imperiale*, Roma, Istituto poligrafico e zecca dello stato, Libreria dello stato.

JEFFERY Lilian H., 1989, *The local scripts of archaic Greece. A study of the origin of the Greek alphabet and its development from the eighth to the fifth centuries B.C.*, revised edition with supplement by A.W. Johnston, Oxford, Clarendon Press, (2003).

IGASMG III = Renato ARENA (dir.), 1994, *Iscrizioni greche arcaiche di Sicilia e Magna Grecia*, Pisa, Nistri Lischi.

IGASMG V = Renato ARENA (dir.), 1999, *Iscrizioni di Taranto, Locri Epizefiri, Velia e Siracusa*, Alessandria, Ed. Dell'Orso.

IGDS I = Laurent DUBOIS (dir.), 1989, *Inscriptions grecques dialectales de Sicile : contribution à l'étude du vocabulaire grec colonial*, Rome, École Française de Rome.

IGDS II = Laurent DUBOIS (dir.), 2008, *Inscriptions grecques dialectales de Sicile. Tome II*, Genève, Droz.

LIDDELL Henry George and SCOTT Robert, 1996[9], *A Greek-English Lexicon,* Revised and augmented throughout by Henry Stuart Jones [...], Oxford, Clarendon Press.

MORPURGO DAVIES Anna, 1987, « Folk-linguistics and the Greek word », dans G. Cardona, et N. H. Zide (dir.), *Festschrift für Henry Hoenigswald. on the occasion of his seventieth birthday*, Ars linguistica 15, Tübingen, Narr, 1987, p. 263-280.

SMYTH Herbert W., 1920, A *Greek Grammar for Colleges*, Cambridge [Mass.], Harvard University Press.

WACHTER Rudolf, 1999, « Evidence for phrase structure analysis in some archaic Greek inscriptions », dans A. C. Cassio (dir.), *Katà diálekton. Atti del III colloquio internazionale di dialettologia greca. Napoli - Fiaiano d'Ischia, 25-28 settembre 1996*, Napoli, Istituto Universitario Orientale (= AION Sezione filologico-letteraria XIX, 1997), p. 365-382.