Teaching Oedipus Tyrannos *with a rich digital edition, in a pandemic*[*]

Introduction

For the Spring Semester (January-May) of 2020, Christopher Blackwell, one of the authors of the present paper, was scheduled to teach an upper-level, undergraduate seminar in Ancient Greek. He had chosen Sophocles' *Oedipus Tyrannos* (henceforth, *OT*) as the text; the goal of the class was simply to have the students read as much of the play as possible in Greek and discuss what they had read.

When he had planned the course and chosen the text, he obviously had no way to foresee that 'Plague, Pollution, and Politics' were going to be quite as topical as they turned out to be in the spring of 2020. However, the question of how to provide a truly *meaningful* reading experience of this masterpiece for students who may never take another course in reading Greek, had engaged both authors for many years. Indeed, for many of us who work with language and textual technologies in the field of Classics, the question is a crucial one. For technologies encourage us to

[*] This article is the product of close collaboration between the two authors. The section entitled *The Setting* was mainly written by C. Blackwell, the section entitled *A Treebank of Sophocles* by F. Mambrini. All the other parts are attributable to both authors in equal measure. The order of the authors is arbitrary.

«FuturoClassico» n. 7, 2021 ISSN: 2465-0951
pp. 18-51

constantly rethink and, in many ways, to reinvent and redefine our approach to 'reading', 'interpreting' or 'learning'.

Thus, when preparing the class, and before the world and the Furman community were affected by COVID-19, Blackwell begun to put his ideas about a 'meaningful reading experience' and about a digital edition that would be at the service of his students into practice. Using some of the technologies that he helped develop, and building on the work and on the friendship of the other author, Francesco Mambrini, he started developing what become his 'zero-fraction edition' of the *OT*. Although the COVID-19 pandemic did not play any role in the process that led to the building of this edition, which largely predates (in some cases, even by many years) the outburst of the disease, it dramatically affected the way our ideas and our tools were put to test in the teaching environment for which they were prepared. In particular, COVID-19, with the ensuing consequences of lockdowns, school closures and solitary study that we are all too familiar with, made the problem of the different types of 'friction' that our edition was designed to solve particularly urgent.

As we will see, with that concept of 'friction' we primarily refer to the time- and labor-consuming interaction between users and tools (be they a paper dictionary, or a web-app in a remote server), whenever students need to 'look up' definitions or brush up even elementary facts of grammars that they simply do not recall. But even leaving all technicality in client-server negotiation aside, an even more general 'friction' that digital resources *must*, in our opinion, deal with is the access bar that prevents most students, *i.e.* those students who do not already possess an extensive training in Ancient Greek language, from reading the work of Greek literature in their original and unabridged form. A tragic and disruptive event like the spread of the COVID-19 exacerbated all these 'frictions', forcing students to isolate themselves (at least physically) from the class, blocking access to printed resources of libraries, and forcing them to rely only on technologies that require a non-trivial amount of resources (especially in terms of bandwidth).

The present article reports on the experience to create one such digital edition that aims at tackling the issues mentioned above. The rest of the introduction describes the context of this pedagogical experience (the teaching and learning environment at the University, the students in the Classics department), as well as the technologies and the background of our work. The second part offers a more detailed overview of the digital edition of the *OT* and offers some conclusions and some 'lessons learned'.

1. *The Setting*

As said, this work stems from a years-long friendship and collaboration between the two authors. The general interests in developing digital resources for the study of ancient languages, and the engagement in a wide network of collaboration that involve scholars in the USA, Germany, Italy, and many other countries of the world, both informally and with a series of short-term projects, have contributed to draw us together. We believe that our experience testifies of the power of digital technologies to cross barriers of language and culture, in our study of the Graeco-Roman heritage, independent of our different, linguistic, educational or professional backgrounds[1].

In this spirit, the paper focuses on a teaching experiment that was carried out in a North American college, namely Furman University. The peculiar nature of this settings is obviously very relevant to the outcome that we report here. In the following paragraphs, we provide some background information that aims to clarify a few relevant aspects of the socio-economic settings of our experiment. This section intends to help readers to understand at least some of the constraints and the opportunities that we had to

[1] Gregory Crane has given many contributions on how digital technologies can empower a truly multicultural and open debate in philological studies, allowing traditions that are outside the club of established powers within the Classics in the Anglo-Saxon world or in Europe to have their voice and their experience heard. See for instance G. Crane, *Greek, Latin and a Global Dialogue among Civilizations*, published online at <https://chs.harvard.edu/gregory-crane-greek-latin-and-a-global-dialogue-among-civilizations/>.

take into account. Hopefully, it should also help readers from other countries to relate to our experience, or to see it in the right context. The intent is not to suggest that what goes on in a North American institution is automatically suitable for the rest of the world. Quite the contrary, we insist on the peculiarities of the settings we worked in in order to provide all data, so that the discussion in an international stage can be informed and stimulating.

Furman University[2], Blackwell's institution, is an example of that almost uniquely American phenomenon of the 'Liberal Arts College', a four-year, all-undergraduate institution offering a general education. Every student must choose one or two specialities to pursue to an advanced level, but all students take a range of courses across the natural and social sciences, the humanities, and the arts. Probably to a fault, the College de-emphasizes training for any particular career, so students studying Chemistry may go on to Law School, students studying Philosophy may end up as carpenters or physicians.

In this environment, a Department of Classics must be thoughtful about its goals and methods. Classics at Furman *has* sent students off to Doctoral programs over the years, and a few alumni have joined the professoriat. But a pedagogy that sees an undergraduate Humanities department as mainly pre-professional training for future Assistant Professors of that department's subject is clearly inappropriate for Furman and its students.

Furman Classics — out of self-interest, and also because its faculty think it is the right thing to do — tries to invite the widest possible range of students to the study of Latin and Greek[3]. In a given year, over the past ten years, an average of 24 students out

[2] It is "Furman University" and not "Furman College" because, around 1900, the Law School for South Carolina was briefly a part of the institution, and the institution has been reluctant to surrender a title that seems to many more exalted. In 2021, the "University" and the undergraduate College are entirely co-terminous.

[3] In this spirit, for instance, one of the tenets in the Department's formal Mission Statement is: «We want as many C students as A students, and we want to serve them equally well».

of a student-body of 2,400 are studying Ancient Greek at Furman, 1% of the student body.

Let it not be forgotten that the full cost of one year at Furman is around $60,000. Let it also be said that a plurality of students receive financial assistance from the school itself or from the government, ranging from a few thousand dollars to the entire cost. Student athletes, in particular, generally get financial assistance, but of course they pay for it in many ways.

Virtually none of these students will go on to graduate school in Classics; very few will go on to graduate school in any Humanist discipline. Most students who study Greek will take two, three, or at most four semesters. This, too, has pedagogical implications. A 19th Century model of Classical language pedagogy that assumes that a student will suffer several years of brute-force memorization, paradigms, and vocabulary-building before being allowed to venture into Xenophon or Caesar, is inappropriate. If a student will take only two semesters of Greek, she deserves to have a meaningful experience, one of which she can boast to future employers. «I completed two thirds of an elementary Greek textbook, ending with the forms of the Aorist and Future Passive» is not a meaningful experience, nor a compelling story.

It is also atavistic, assuming a world, long past, when 'looking things up' was a laborious, mechanical process that would necessarily cause anything that seemed like 'reading' to grind to a halt. At least in the English-speaking world, that world obsolesced in the 1990s, when the visionary Classicist Gregory Crane oversaw the transfer of The Perseus Project from distribution on CD-ROM to availability on the World Wide Web, *and* made the Promethean decision to make it freely available for everyone in the world with an internet connection[4]. While specialists would still need to 'look

[4] *The Perseus Digital Library*, Gregory Crane, Ed. Tufts University, <http://www.perseus.tufts.edu>. In 2021, other web and mobile applications offer services of "dictionary lookup" that link dictionary entries to words in digital editions of Greek and Latin texts. In particular, we wish to mention *Logeion*, hosted at the University of Chicago, which allows users to browse

up' many things on many aspects of the language, several of which can also be dug from many excellent scholarly resources online, dictionary entries for basic vocabulary are, thanks to initiatives like Perseus, readily, quickly and openly accessible.

2. *The Data*

The authors of the present paper have been privileged to play a part in the revolution sparked by Crane's vision. Both have worked to address the challenges of documenting ancient texts and integrating as much data as possible in the service of readers.

Indeed, each of us has taken an active role in what are perhaps two of the most exciting and foundational acquisitions in the Digital Classics communities of the last decade. Both digital resources, the morpho-syntactic annotated corpora known as 'treebanks', and the CITE architecture, played a major role in the enhanced edition of the *OT*. The following paragraphs introduce both and illustrate how they are used to shape the data that make the backbone of our digital edition.

3. *A Treebank of Sophocles*

During a previous iteration of a course dedicated to reading Sophocles in Greek at Furman, several students discovered and became fascinated with Mambrini's work documenting the syntax of the *OT* via syntactic dependency treebanks.

Treebanks are annotated corpora that store sentence-by-sentence and word-by-word annotation about morphological aspects, starting from the part of speech, to (in several case) lemmatization and full description of all the features in word flection (number, tense, mood, gender, degree, etc.). Moreover, treebanks also implement some formalism to account for the syntactic relations and functions of every word in each sentence. To be more specific, the

multiple dictionaries including, alongside English, German, Dutch and French lexicons. The pioneer role of the Perseus DL in this respect, however, is undisputable.

minimal units of a treebank that receive both morphological and syntactic annotation are called 'tokens' (and treebank tokens include punctuation marks!); although tokens usually correspond to our ordinary notion of words, they do not necessarily coincide with them[5]. Since the creation of the first projects in the 1990s, treebanks have become widely used for corpus-based linguistic research and as training data for Natural Language Processing applications[6].

In 2008, G. Crane and the Perseus Project launched the *Ancient Greek Dependency Treebank*, a comprehensive treebank of Greek literary texts that complements the *Latin Dependency Treebank*, started in 2006[7]. (Together, the treebanks are referred to as the *Ancient Greek and Latin Dependency Treebank*, or AGLDT). Mambrini curated the annotation of the seven fully preserved tragedies traditionally attributed to Aeschylus, which were published in 2009, and of four tragedies of Sophocles published in 2011[8]. In the

[5] In fact, it is important to note that the basic unit of treebank annotation is the *syntactic* unit. This means that, for instance, in the case of negative conjunctions like οὔτε or μήτε – and the same would be true for English 'neither' – the two units that serve different syntactic functions (the negative particle and the coordinating conjunction) are split into two tokens. Similarly, clitics that are written together with a lexical word (like Latin *-que*) or that are fused with it on account of phonetic phenomena such as the crasis can be also split into two tokens.

[6] For an introduction to treebanks and their applications, see the essays collected in A. Abeillé (ed. by), *Treebanks. Building and Using Parsed Corpora,* Kluwer, Dordrecht-Boston 2003. On treebanks for Greek and Latin see also F. Mambrini, *L'Ancient Greek Dependency Treebank. Un Nuovo Strumento per Lo Studio Della Lingua Greca*, «Lexis» 29, 2011, pp. 51-70.

[7] See D. Bamman *et al.*, *An Ownership Model of Annotation: The Ancient Greek Dependency Treebank*, in M. Passarotti *et al.* (ed. by), *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, EDUCatt, Milan 2009, pp. 5-15. D. Bamman-G. Crane, *The Ancient Greek and Latin Dependency Treebanks*, in C. Sporleder *et al.* (ed. by), *Language Technology for Cultural Heritage. Theory and Applications of Natural Language Processing*, Springer, Berlin-Heidelberg 2011, pp. 79-98.

[8] Mambrini annotated Sophocles, *Trachiniae, Antigone, Oedipus Tyrannos,* and *Electra.* The *Ajax* was annotated by D. Libatique; *Philoctetes* and *Oedipus at Colonus* are still not annotated.

following years, the collection was enriched with several texts, and it sparked the creation of several parallel projects of treebank publications that adopt the same annotation format[9].

To encode information about the syntactical structure of a sentence, the AGLDT uses a formalism that is based on a theory of syntax known as 'Dependency Grammar'[10]. In particular, the AGLDT is based (with some minor modifications) on the same guidelines for dependency annotation as the one used in the *Prague Dependency Treebank* (PDT) of Czech[11]. Contrary to phrase-structure formalisms, a dependency grammar does not describe the syntax of a sentence in terms of relations between constituents or phrases (such as 'Noun Phrases' or 'Verb Phrases'), but establish relations directly from a dependent to its governing head. These relations are acyclic: a word cannot depend on itself, not even indirectly through one of its descendants; also, a word cannot have more than one head. Accordingly, the sentence structure that results can be visualized as a tree-shaped graph, with the tokens serving as the nodes.

[9] The AGLDT data is available at <https://perseusdl.github.io/treebank_data/>. Currently, the full ecosystem of the Pereseus treebanks assemble annotations for a total close to 1M words, including the full text of *Iliad* and *Odyssey*, and many prose texts of the Classical and Hellenistic eras annotated by Vanessa Gorman. For an overview and some quantitative data see A. Keersmaekers *et al.*, *Creating, Enriching and Valorizing Treebanks of Ancient Greek*, in M. Candito *et al.* (ed. by), *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, Association for Computational Linguistics, Paris 2019, pp. 109-117.

[10] 'Dependency grammars' are a family of theory of syntax that have a long history and are articulated in a variety of different versions. Modern dependency theories generally refer to L. Tesnière, *Éléments de syntaxe structurale*, Klinksieck, Paris 1959, as a precursor. See also T. Osborne, *A Dependency Grammar of English: An Introduction and Beyond*, John Benjamins, Amsterdam 2019, for an introduction.

[11] In its turn, the PDT is based on a theoretical framework, developed at Prague, known as Functional Generative Description. See P. Sgall *et al.*, *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*, Academia, Dodrech 1986, and Mambrini, *L'Ancient Greek Dependency Treebank* cit.

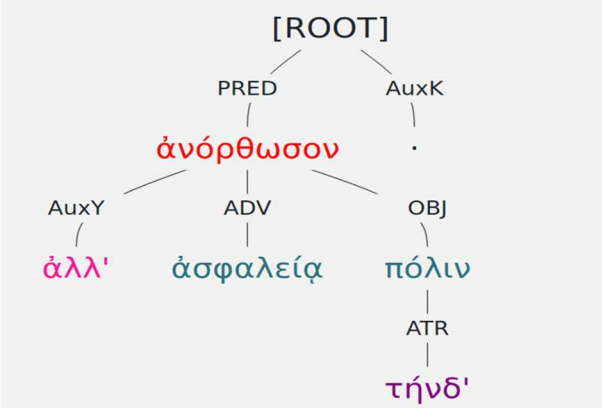

Fig. 1: *Sophocles,* OT *51: a dependency treebank.*

Figure 1 represents the syntactic tree of one sentence from the *OT* (line 51). As it can be seen, the tree, like all prototypical sentences in dependency treebanks, is rooted on the main predicate (the imperative ἀνόρθωσον, in this case), with the exception of the sentence-ending punctuation mark, and the verb governs the direct object (πόλιν). The noun, in its turn, is the head of the demonstrative τήνδε. Note, also, that all tokens, including particles and adverbials like ἀλλά, and all punctuation marks receive an annotation, so that no part of the text is left without metadata.

Treebanks are extremely useful to encode, visualize and query the syntactic structure of a text, as well as all the interpretative aspects that can be linked to the morphology and the syntax of a passage[12]. Although a formalism like the AGLDT is incapable of conveying information about a range of questions that are crucial for the interpretation of a tragedy (such as the illocutionary force of utterances or the sense of words), the number of issues about which treebank annotators *are* able to encode their interpretation

[12] For an example of a treebank-based investigation on one syntactic question in Ancient Greek (the competing agreement patterns with coordinated subjects) see F. Mambrini-M. Passarotti, *Subject-Verb Agreement with Coordinated Subjects in Ancient Greek*, «Journal of Greek Linguistics» 14, 2016, pp. 87-116.

in the annotation is quite wide[13]. Ultimately, treebankers are forced to take a decision and opt for one reconstruction over the other potential readings, even in case of incertitude or actual ambiguity that are certainly not infrequent in the extant tragedies. However, tools and editions that interact or are based on treebank data should always allow teachers to emphasize the annotation as a *process*, rather than as a finished product. Treebank annotation captures and asserts *one reading* of a sentence, and makes it explicit, and thus subject to analysis, criticism, or indeed refutation. We believe that the technologies should put teachers in the position, if they want, to engage students in questioning the morpho-syntactic reconstruction and investigate both the ratio and possible alternatives to the interpretation that is written into a file[14].

4. *The CITE Architecture and the "CITEification" of the Treebank*

In that previous seminar where the annotation on the *OT* was introduced to them, the Furman students of Blackwell extended the treebank with a few additions for their final project; in the process the XML format, in which the AGLDT files are distributed, was converted into a tabular structure. The result was a tabular dataset, with one row for each word in the play[15].

The rows of the table were structured so that each one represented a token in the play in sequence: the first row was for the word «ὦ»; the second for «τέκνα», *etc.* The columns represented a series of features that reflect three different layers of information that the students attempted to capture. The first and more obvious

[13] For a discussion and an example see F. Mambrini, *The Ancient Greek Dependency Treebank: Linguistic Annotation in a Teaching Environment,* in G. Bodard and M. Romanello (ed. by), *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement*, Ubiquity Press, London 2016, pp. 83-99.

[14] The point is again vigorously argued in Mambrini *The Ancient Greek Dependency Treebank: Linguistic Annotation* cit. See also below, n. 35.

[15] This dataset resided as a Google Fusion Table until December of 2019, when Blackwell retrieved it and archived it as a CSV file on GitHub: <https://github.com/Eumaeus/Oedipus_2019/tree/master/data/ot_all.csv>.

is the text itself. The second level captures the grammatical information stored in the treebank for each token/row: the lemma (τέκνον for the second token, Κάδμος for the third, *etc.*), the full morphological description (noun neuter vocative plural and noun masculine genitive singular, respectively), and the labeled syntactic relation of a token to its head.

There is however a more fundamental type of information that is attached to tokens, whose importance may escape the notice of everyday users, but is nonetheless essential for a digital architecture. This layer ensures recovery and identification of data points in themselves and within their textual context, both for humans and machines. Human readers – we may add: *trained* human readers, for a *modicum* of training is needed to acquire familiarity with citation habits – would identify the second token of the play using an expression like: «τέκνα, in line 1 (of Sophocles *OT*)». Greek tragedy, like much poetry, is indeed traditionally cited by line-number, with for instance citation strings like: 'Sophocles, *OT* 123-125.' This concise and precise way to encode the relevant information has served scholarship well for centuries.

The concept of 'canonical citation' is the basis of CITE, a digital library architecture originally developed for the Homer Multitext and subsequently employed for other digital humanities projects[16]. CITE provides an architecture to identify, retrieve, and manipulate data by means of machine-readable and machine-actionable identifiers based on canonical citations[17].

For reproducible digital analysis, canonical citation is vital; we must be able to identify the text under analysis, with whatever precision is required. A digital presentation of a text is simply one form of analysis, as is 'linking' individual words to morphological and lexical information.

[16] On the Homer Multitext project see most recently C. Dué-M. Ebbott, *The Homer Multitext within the History of Access to Homeric Epic*, in M. Berti (ed. by), *Digital Classical Philology*, De Gruyter-Saur, Berlin-Boston 2019, pp. 239-256.

[17] On the CITE Architecture see C. Blackwell-N. Smith, *The CITE Architecture: a Conceptual and Practical Overview*, in Berti, *Digital Classical Philology* cit., pp. 73-94.

Both in the ordinary use of canonical citations and within the CITE model, references are structured hierarchically: to quote one passage of the *Iliad* (*e.g.* I, 10), we are trained to employ a tuple of references, constructed starting from the larger unit (the book, *e.g.* 'I') followed by the line number ('10').

A CTS-URN, the identifier used in the CITE architecture and in the 'Canonical Text Services' for text passages, captures this hierarchy, based on the model of an «ordered hierarchy of citable objects» (OHCO2)[18]. Examples of CTS-URNs are such strings as:

urn:cts:greekLit:tlg0012.tlg001:1.26

where, to simplify the matter slightly, the segments 'tlg0012' and 'tlg001' represent respectively the textgroup (the 'author') and the work (Homer and the *Iliad*, in this case), while the segment '1.26' reflects the hierarchy of the passage reference (book I, line 26)[19].

The CITE Architecture thus proved a suitable framework for the students to represent all layers of information, and to provide all elements of the text with stable identifiers. In the final version of the tabular serialization, the columns of each row, representing all annotated tokens in the treebank, recorded the following information:

- Line number.
- Sequence number.
- A CTS-URN uniquely identifying the word.
- A CITE2-URN identifying the sentence of which the word is a part.
- A sequence number for the sentence.
- A CITE2-URN uniquely identifying this word as an "analyzed token".

[18] OHCO2 is a cross-platform library for working with citable data. Its documentation can be accessed at <http://cite-architecture.org/ohco2/>.

[19] Readers will find a more detailed explanation of the CTS-URNs and their syntax at <http://cite-architecture.org>; see in particular C. Dué *et al.*, *A Gentle Introduction to CTS & CITE URNs*, in Id. *Homer Multitext Project: documentation*, published online at <https://www.homermultitext.org/hmt-doc/guides/urn-gentle-intro.html>, 2012.

- Fields capturing the word's morphological characteristics, as documented in Mambrini's treebank.
- The CITE2-URN of the token of which this word is a dependent, based on Mambrini's treebank.
- A CITE2-URN identifying the nature of the syntactic relation between this word and its "head".
- Which character in the play is speaking.
- A CITE2-URN to the entry for this word in the online Liddel-Scott-Jones *Greek-English Lexicon.*

Instead of simply providing a string for the lemma of each inflected forms found in the text (*e.g.* ἀνορθόω for the aorist imperative ἀνόρθωσον of *OT*, 51), this tabular version links the token with the URN of the appropriate dictionary entry in a digital version of the Liddel-Scott-Jones dictionary[20]. The benefit of such 'Linked-Data' approach to lemmatization are manifold, but can be readily summarized under the rubric of less ambiguity and increased interoperability. On the one hand, the use of a stable identifier allows for more precise reference in case of ambiguity (to identify, for instance, which of the two verbs δέω registered in the *LSJ* we are lemmatizing a given form under). On the other hand, if other digital resources on the WWW point to an entry or add information about it, then this new information becomes discoverable and usable also for our tokens that are lemmatized with the *LSJ* URN[21].

[20] The Liddel-Scott-Jones *Greek English Lexicon* was originally digitized for inclusion in the Perseus Digital Library; see J.A. Rydberg-Cox, *Mining Data from an Electronic Greek Lexicon*, «Classical Journal» XCVIII, 2, 2002, pp. 183-188. A revised version is also offered on the *Perseus Under Philologic* website, hosted at the University of Chicago: <http://perseus.uchicago.edu/>. The Furman CITE-compliant version was derived from these two original editions. For more information, see <http://eumaeus.github.io/2018/10/30/lsj.html>, and the follow-up discussions at <http://eumaeus.github.io/2018/11/04/lexService.html> and <http://eumaeus.github.io/2018/11/05/chicago.html>.

[21] On the benefits of using stable identifiers (and the other best practices of the Linked Open Data paradigm) for lemmaization, see M. Passarotti *et al.*, *Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin*, «Studi e Saggi Linguistici» 58, 2020, pp. 177-212.

The dataset is immediately revealing to anyone planning to read this play with students who have studied at most three semesters of Ancient Greek. With the caveat that, as we said, the 'tokens' include punctuation, we can instantly discover some important facts on the lexicon and the morphology of the play. In total, there are 11,195 tokens (words and punctuation) in the *OT*, with 3,874 distinct tokens (i.e individual forms of words, and discrete punctuation marks) and 1,842 distinct *lemmata*. Of these 1,842 distinct *lemmata*, 1,306 are represented only once or twice in the play; 1,021 are represented only once. The verb εἰμί is present in 47 forms (accounting for differences in accentuation of the enclitic forms). The list of other verbs, ranked according to the number of morphological forms appearing in the play, is what one might predict who knows the plot: λέγω (21), ἔχω (19), δράω (19), φαίνω (18), οἶδα (18), κτείνω (14), ἀκούω (14), ὄλλυμι (14)[22].

This data looked like an opportunity. For many of the students about to read the *OT*, this would be their last encounter with Ancient Greek, perhaps for the rest of their lives. Some of these

As an instance of a digital project that makes use of the URNs of the Furman *LSJ*, readers can be referred to the digitized *IGVLL* published by the LiLa project, on which see Franzini *et al.*, *Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin*, in J. Monti *et al.* (ed. by), *Proceedings of the Seventh Italian Conference on Computational Linguistics*, CEUR-WS.org, Bologna 2020, published online at <http://ceur-ws.org/Vol-2769/paper_06.pdf>.

[22] The students who originally compiled this data discovered other interesting things, such as that while both surface-forms and *lemmata* follow a distribution-curve that supports Zipf's law, other things do, too, like the frequency of verbal tenses. 'Zipf's law' is an empirical relation that holds between rank and frequency and is often observed, among other areas, in many facts of language, most notably lexical distribution. Users of language corpora can verify that, in accordance with the 'law', there is an inverse relation between the frequency rank of a word and the number of its occurrences, so that the most frequent word occurs about twice more often than the second, about three times more often than the third, and so forth. See G.K. Zipf, *Selected studies of the principle of relative frequency in language*, Harvard University Press, Cambridge, MA 1932. J. Van de Walle-K. Willems, *Zipf, George Kingsley (1902-1950)*, in K. Brown (ed. by), *Encyclopedia of Language & Linguistics*, 2nd ed., Elsevier, Boston 2006, pp. 756-757.

students did not come to their undergraduate studies with sophisticated experience in language and literature; several were athletes, who always have many demands on their time and energy (demands that they *must meet* in order to continue their education).

5. *Data as 'Commentary'*

In previous years, we have taught with various commentaries, some aimed at a professional scholarly audience, and some 'student commentaries'. In every case, a commentary is necessarily selective, and it often seems that even the most thoughtful ones frequently assume knowledge a given student simply does not have, or had six months earlier but has forgotten. Any commentary that says, *e.g.*, «Common words assumed to be in the reader's vocabulary are not included in the Vocabulary» is only going to discourage a student who has forgotten the distinction between κελεύω and καλέω, or the distinction between ἴθι and ἴσθι.

For the sake of argument we might assume that a very gifted, fourth-semester student holds in her active memory the forms and the meaning of every word that occurs at least twice in the play. Based on the numbers we reported above, though, even this optimistic expectation still condemns our gifted student to the time-consuming prospect of looking up 1,021 words and recognizing their forms, *before* she can even think about syntax, let alone meaning and artistry.

Of course, every rational student will keep a translation at hand. Perhaps some American professors of Greek still command their students to avoid looking at translation — if they do, it is absolutely in vain — but at Furman we encourage students to consult as many translations as they like. A translation is, after all, essentially a comprehensive commentary on the text in a very efficient package. Anything in service of getting us to the point where we can have an informed and confident conversation about the Greek text!

A translation is one way to *capture and serialize* a *reading* of a text at a granular level. A syntactic treebank, however, is another, that operates at an even a more granular level.

With the treebank data we have a serialization of (one) reading of Sophocles' *Oedipus Tyrannos*; it is a comprehensive commentary that is directly aligned, word-by-word and sentence-by-sentence with the Greek. With the available *OT* treebank, based on the work of Mambrini, Furman students would enjoy a commentary by an annotator who, as we saw, did not elide common words or form, assuming they were known. Every word is parsed, disambiguated, and fitted into its proper (as asserted by the Editor) slot in a syntactic graph.

6. *A Rich, Zero-Friction Digital Edition. Prolegomenon to User Interface*

In the weeks before the start of Spring Semester 2020 at Furman, Blackwell began to experiment with what he thought of as a 'zero-friction reading environment.' As we mentioned, the 'friction' that he wanted to remove, or at least minimize, is the inevitable 'looking up' that stands between readers of Greek and their immediate understanding of the lexicon of a text.

The inspiration for this idea was the brilliant work of Professor Randall Childree, also working at Furman University, who has thought deeply about language pedagogy, and regularly prepares for his students online reading environments[23]. The features we admire about Childree's work, apart from the philological rigor, include the Aldine elegance of the design and attention to typography and color, and the compact presentation of morphological and lexical information. But mostly, it is the *speed and efficiency* with which that information is delivered that make Childree's student editions particularly well suited for their intended audience.

What makes Childree's editions so efficient is the fact that their creator has embedded morphology and lexicography into the

[23] One example, with the text of the *Life of Caligula* from Svetonius, *The Lives of the Twelve Caesars*, can be browsed at <https://rchildree.github.io/suetonius-caligula/>.

HTML of the page, using such common web technologies as JavaScript and CSS, to deliver it for any given word upon a *hover* event (that is, when the mouse cursor is positioned over a text passage). As modern computers become ever faster, and every modern browser's JavaScript engine is every more optimized, there is no discernable delay between (a) pointing at a word, and (b) learning about the word.

Because Childree's HTML pages for readers embed all necessary data they are responsive and also much more, and much more sensibly, efficient. While web services like the Greek and Latin Word Study Tool of the Perseus Digital Library, which work on servers where people can send their requests with forms to analyze[24], will always be the most efficient solution for countless users in many situations, Childree's approach answers one fundamental objection. Given that pre-parsed texts are already available (*e.g.* in treebanks) or can be produced either manually or with the help of the existing software, is it really necessary to send a request to a complex parsing application on a distant server every time a student cannot remember that *pater* means 'father'?

Blackwell's motivation, in December, 2019, was to emulate Childree's model of user-interface for readers while delivering all of the rich data offered by Mambrini's comprehensive morpho-syntactic analysis of the play.

7. *Data Modeling and the CITE Architecture*

The starting point for our work, and for any 'Oedipus Reading Environment' to come, is a digital edition of the text, formatted in CEX format. CEX, which stands for CITE Exchange Format, is a plain-text, line-oriented data format for serializing citable content

[24] Users may try the Word Study Tool at <http://www.perseus.tufts.edu/hopper/morph>. On the Morpheus analyzer, that powers up the service, see G. Crane, *Generating and Parsing Classical Greek*, «Literary and Linguistic Computing» VI, 4, 1991, pp. 243-245.

following the models of the CITE Architecture, as discussed above[25]. An excerpt of a text expressed as CEX looks like this:

```
#!ctsdata
urn:cts:greekLit:tlg0011.tlg004.fu:1#ὦ τέκνα, Κάδμου τοῦ πάλαι νέα τροφή,
urn:cts:greekLit:tlg0011.tlg004.fu:2#τίνας ποθ' ἕδρας τάσδε μοι θοάζετε
urn:cts:greekLit:tlg0011.tlg004.fu:3#ἱκτηρίοις κλάδοισιν ἐξεστεμμένοι;
urn:cts:greekLit:tlg0011.tlg004.fu:4#πόλις δ' ὁμοῦ μὲν θυμιαμάτων γέμει.
```

As it is intended for the serialization of citable content, CEX is also able to capture the hierarchical model of citations (OHCO2) that we mentioned before when we discussed the structure of CTS-URNs. The string that is associated to each line, and serves as a unique persistent identifier to it is a CTS-URN that captures the citation hierarchy, which, in this case, is very simple, and only one-level deep: poetic lines are the only unit identified, following a schema which is both canonical and traditional. Every passage of text is part of a citation-object (URN + text), and the set of citation-objects represents the entirety of the text. A version of the text that conform to this model is ideal for many kinds of computational analysis because there is nothing present except the language of Sophocles' play — no internal notes, comments, or other markup that a process would have to discard or work around.

The text above is one *version* of the play, the most fundamental instantiation of the abstract notion of the *Oedipus Tyrannos*, and provides a citable edition of the work. That Edition would not be sufficient, however, as the basis for a rich reading-environment. To align morphological, lexical, and syntactic data to the text for the benefit of readers, it is necessary to identify each token in the text, explicitly and unambiguously. For even a simple action like isolating the units of analysis, a preprocessing task that is known in computational linguistics as 'tokenization', rests on method-

[25] CEX specifications can be found at <https://cite-architecture.github.io/citedx/CEX-spec-3.0.1/>.

ological assumptions and is, in a sense, an act of interpretation[26]. Rather than relying on string-manipulation, the CITE Architecture allows for the creation and cross-referencing of 'analytical exemplars', versions of a text, derived from an Edition and therefore explicitly aligned with it. One commonly used type of Analytical Exemplar is a tokenized exemplar. One advantage of this approach is that, in contrast with many forms to tokenization that depend on XML markup, CITE and analytical exemplars allow scholars to work with many different kinds of tokenization of the same text. A text might be tokenized by words and punctuation, as for syntactic analysis, or just words (removing punctuation), or by metrical foot, or by syllable, and many types of tokenization can coexist together with the text.

For the *OT*, a fragment of such a tokenized Exemplar, in CEX format, would look like this:

```
#!ctsdata
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.1#ὦ
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.2#τέκνα
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.3#,
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.4#Κάδμου
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.5#τοῦ
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.6#πάλαι
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.7#νέα
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.8#τροφή
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1.9#,
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:2.1#τίνας
```

[26] As it was already said, treebank tokenization, which relies on the notion of syntactic unit, is but *one* type of tokenization; in some cases, it may produce results that differ from the ordinary notion of typographic words. It goes also without saying that another form of tokenization required by treebanks, the 'sentence splitting', *i.e.* the identification of sentences and sentence boundaries, is a task that even editors of printed critical editions must pay attention to. In some manuscripts of Greek tragedies, for instance, we find cases where sentence-ending punctuation is placed where the position of a particle (like γάρ in Aeschylus, *Agamemnon*, 222 or Sophocles, *Electra*, 492) was felt as wrong; these interventions result in sentence divisions that are generally deemed unacceptable by modern editors. See E. Fraenkel, *Aeschylus. Agamemnon*, Clarendon Press, Oxford, 3 voll.: vol. II, *Commentary on 1-1055*, 1950, p. 128.

urn:cts:greekLit:tlg0011.tlg004.fu.tokens:2.2#ποθ'
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:2.3#ἕδρας
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:2.4#τάσδε
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:2.5#μοι
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:2.6#θοάζετε
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:3.1#ἱκτηρίοις
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:3.2#κλάδοισιν
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:3.3#ἐξεστεμμένοι
urn:cts:greekLit:tlg0011.tlg004.fu.tokens:3.4#.

The citations here now capture a two-level deep hierarchy: poetic line + token sequence (within the line). The top-level identifier matches those in the Edition (lines 1-3); the second-level identifier is unique to this exemplar. The 'bibliographic hierarchy' part of the CTS URN – urn:cts:greekLit:tlg0011.tlg004.fu.tokens: – makes explicit that this exemplar, '.tokens', is derived from and aligned to the 'Edition, urn:cts:greekLit:tlg0011.tlg004.fu:'.

A request for the text of 'urn:cts:greekLit:tlg0011.tlg004.fu.tokens:1' will deliver the same text-content as 'urn:cts:greekLit:tlg0011.tlg004.fu:1', in the form of 8 citation-objects, rather than one.

Up to this point, we have described an Edition of the play and a derived Exemplar. With other projects in creating digital texts, we would have started with a digital version of the work, such as the TEI-XML editions in the very large library made available by the Perseus Project[27]. The work then would be to tokenize that Edition, and then align the resulting Exemplar with the syntactical analysis and its attendant data found in the treebank.

That process would be fraught and error-prone, requiring meticulous checking of every step. It would also be completely unnecessary, since the treebank data provide a complete text of the play when the tokens are read sequentially. So the Edition- and

[27] The PerseusDL repository of openly-licensed XML texts is available at <https://github.com/PerseusDL/canonical>. TEI stands of 'Text Encoding Initiative' and is a widely used standard to produce digital curated editions. For a highly readable introduction to TEI, see L. Burnard, *What is the Text Encoding Initiative?*, New online edition, OpenEdition Press, Marseille 2014, freely available at <http://books.openedition.org/oep/426>.

Exemplar-building was a matter of 'reverse engineering' the *OT* syntax into different versions of the text of the play, as shown above.

One additional analytical process was required in order to present this data as a coherent text for human readers: speaker attribution.

The identification of the 'Speaker' is an interesting case of relation between text and metadata. This piece of information is both inherent and entirely necessary to the text, but is also external to the text itself[28]. The *OT* is a theatrical play, so it is vital for readers to know who is speaking, and the information is encoded in the reproductions of dramatic texts since our remotest sources (from the *paragraphos* used in antiquity to mark the change of speakers, to the modern practice of printing the name of the characters on the side). At the same time, it is undesirable to have speaker-attributions present when, for example, counting words. It would be misleading to claim that 'Χορός' occurs 46 times in the language of the *OT*, when in fact it does not occur at all. A TEI-compliant digital edition would treat this information as metadata surrounded with special tags, like 'speaker'[29]. In the enhanced dataset from which we started, a speaker-attribution is attached to every individual token. Clearly, however, that would be *too much* information to present to a reader.

The CITE Architecture accommodates this, again, through derived analytical exemplars. In this case, we programmatically constructed an exemplar from the enhanced, tabular, dataset that extended the Edition's citation scheme. Below is an excerpt from the analytical exemplar ('urn:cts:greekLit:tlg0011.tlg004.fu.sp:'). The 'tlg0011.tlg004.fu' indicates once more 'Sophocles, *Oedipus Tyran-*

[28] For another example of how to use CTS-URNs to automatically extract the speaker attribution from Perseus TEI editions and to link them to the treebanks, as well as for a treebank-based study on the syntax of the Sophoclean characters, see F. Mambrini, *The syntax of the heroes? A treebank-based approach to the language of the Sophoclean characters*, to be published in «Classics@».

[29] See the TEI guidelines (v. 5.0) at <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-speaker.html>.

nos, fu-edition'. The '.sp:' extension is the exemplar ID ('speaker-exemplar'). In the data-excerpt below, the CTS-URNs are truncated to show only the *citation field* and the text, for legibility. The CEX-delimiter '#' has been replaced with a tab-character.

```
:1170.speaker1    Οἰδίπους
:1170.1           κἄγωγ' ἀκούειν: ἀλλ' ὅμως ἀκουστέον.
:1171.speaker1    Θεράπων
:1171.1           κείνου γέ τοι δὴ παῖς ἐκλῄζεθ': ἡ δ' ἔσω
:1172.1           κάλλιστ' ἂν εἴποι σὴ γυνὴ τάδ' ὡς ἔχει.
:1173.speaker1    Οἰδίπους
:1173.1           ἦ γὰρ δίδωσιν ἥδε σοι;
:1173.speaker2    Θεράπων
:1173.2           μάλιστ', ἄναξ.
:1174.speaker1    Οἰδίπους
:1174.1           ὡς πρὸς τί χρείας;
:1174.speaker2    Θεράπων
:1174.2           ὡς ἀναλώσαιμί νιν.
:1175.speaker1    Οἰδίπους
:1175.1           τεκοῦσα τλήμων;
:1175.speaker2    Θεράπων
:1175.2           θεσφάτων γ' ὅκνῳ κακῶν.
:1176.speaker1    Οἰδίπους
:1176.1           ποίων;
:1176.speaker2    Θεράπων
:1176.2           κτενεῖν νιν τοὺς τεκόντας ἦν λόγος.
```

For this exemplar, the intended 'analysis' is 'reading by human reader.' Note how the *canonical citation scheme* of the play is preserved. A request for line 1173 would return:

```
:1173.speaker1    Οἰδίπους
:1173.1           ἦ γὰρ δίδωσιν ἥδε σοι;
:1173.speaker2    Θεράπων
:1173.2           μάλιστ', ἄναξ.
```

When several lines are by the same speaker, the speaker-attribution is not repeated. But this also allows for multiple speakers, trading half-lines (*antilabé*). The values 'speaker1 and 'speaker2' are arbitrary identifiers, but they are predictable, and an applica-

tion or transformation can catch those 'passages' and treat them differently[30].

8. *Chunking for Building*

The process that we described up to this section are sufficient to produce an exemplar for a reader, with speaker-attributions, and an aligned tokenized exemplar keyed to syntactic and morphological information. This is almost everything necessary to make a rich edition.

Most of the details of the code libraries that generated the final edition of the play are beyond the scope of this article. One important aspect that readers must keep in mind is that a relatively long text like the *OT*, with the amount of information that we intend to add to it, cannot be served in one single web page. The text must be split into smaller units, or 'chunks'. We briefly list here only the general steps that we implemented (using Scala)[31] to produce an HTML output similar to Childree's student editions:

- Read the Editions and Exemplars from CEX into a CITE-Library object in data.
- Read the enhanced dataset, with morphological and syntactic data into a Vector (list) of data-objects defined by a bespoke Scala class called 'OToken'.
- Chunk the text in some sensible and practical way, for readers.
- For each chunk, write out an HTML page and an accompanying page of Javascript and JSON (for exposing syntax graphs).

The chunking algorithm, that segments the play into sections to be presented on a single page (the chunks) depends on the 'speaker-exemplar'. It accepts number of lines as a target (the default is 25 lines per HTML page), and then uses tail-recursion to read through the speaker-exemplar, gathering lines until the

[30] While CITE and CTS are generic protocols, and the above presentation of the play would 'work' in any generic implementation of CTS, there comes a time when knowledge of the data can be helpful. Another play transformed into an example with 'speaker1' and 'speaker2' citation-values would work with the same code that transformed out *OT*.

[31] See <https://www.scala-lang.org/>.

target-number is reached, and then continuing to the end of the current speech[32].

9. *The HTML Pages*

A great amount of information is written to each page, including morphological identification of each word, a URN to a lexicon-entry, and a short-definition for each word extracted from the CITE-Architecture version of the *LSJ*[33].

In addition — because the enhanced dataset from the treebank includes sentence-by-sentence alignment, each HTML page is matched with a page of Javascript and a JSON expression of the syntax tree for each sentence in the HTML. Most of this data is hidden by default, to be exposed via Javascript and CSS when the user hovers or clicks on element.

10. *The Lexicon Short-Entries*

The Perseus Digital Library innovated in presenting a 'short definition', mechanically extracted from a longer lexicon entry, for a quick reference, when a student has forgotten what καλέω means, for example, but does not need to read the entire article on the verb (and its 11 senses, nested in various level of complexity) in the big lexicon. In the original implementation, Perseus simply grabbed the first English word from the *LSJ* entry. This solution, however, could lead to confusion, if the first non-Greek word were, for example, a Sanskrit cognate, or a descriptive part of the article's prose, like «rarely…». For our edition of the *OT*, we adopted a different approach in which the code libraries grabbed

[32] 'Tail-recursion' is a looping operation that is sparing of system memory. It is useful for loops whose duration is hard to anticipate, and where large amounts of data are going to be read into memory on each loop. The classic example is a factorial-operation, where a traditional 'loop' structure would quickly exhaust memory. Scala's implementation of tail-recursion, at the compiler level, was helpful in building the JSON that captures the syntactic graphs of each sentence.

[33] See above, n. 15.

the entire lexical article, disposed of all non-'foreign' words, and simply truncated the list after a few hundred characters. This generally ensures that the basic meaning of a word will be present, without overwhelming the reader. Those short entries are linked to the full online lexicon (that link is the only feature of the digital edition that requires a network connection).

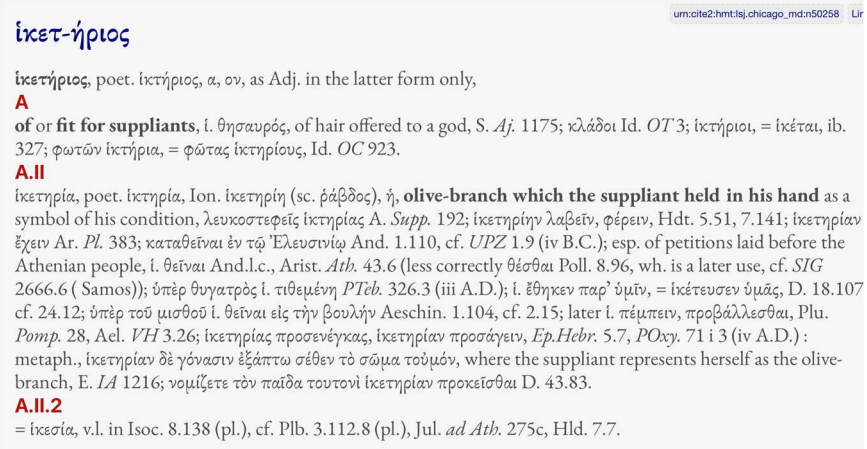

Fig. 2: *Entry for ἱκετήριος in Furman's CITE-compliant LSJ dictionary.*

To take for example the word «ἱκτηρίοις» at *OT* 3, its full *LSJ* entry begins with the words:

> **of** or **fit for suppliants**, ἱ. θησαυρός, of hair offered to a god…

In the original Perseus model, the application would see the first 'English' offered, which unfortunately is 'of', which would be offered as the short-definition to readers. This has caused great confusion over the years!

Our edition of the *OT*, on the other hand, presents the information visualized in fig. 3, when a user hovers over "ἱκτηρίοις" in line 3.

Obviously, the reader has one-click access to the full *LSJ* entry at any time.

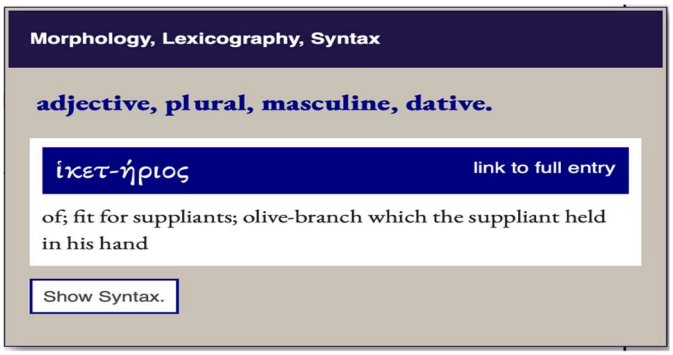

Fig. 3: *Entry for ἱκετήριος in Furman's CITE-compliant LSJ dictionary.*

11. *The User Interface and Reading Experience*

Our online edition of the *Oedipus Tyrannos* is available for everyone at <http://folio2.furman.edu/ot/pages/index.html>. What follows is a short tour of the design decisions, with commentary specific to the fact that, as this edition began to be used for teaching, the course at Furman University went all online. Some elements in the User Interface were modified based on the teacher's and students' experiences in this new environment.

The current build of the site divides the play across 59 pages. When a reader comes to a page, the default view (based on the "speaker-exemplar") shows speakers and poetic lines (or half-lines) in a traditional manner.

Hovering a cursor over a line highlights it. In the initial versions this was not the case, but when sharing a screen, it proved extremely helpful to have this obvious way to draw a class's attention to the line currently under consideration. The default view offers no other interaction until a user clicks the "reveal" triangle next to the line, at which point that line expands to display data from the "tokenized exemplar".

Θεράπων

1176 κτενεῖν νιν τοὺς τεκόντας ἦν λόγος.

Οἰδίπους

1177 πῶς δῆτ' ἀφῆκας τῷ γέροντι τῷδε σύ;

Θεράπων

1178 κατοικτίσας, ὦ δέσποθ', ὡς ἄλλην χθόνα

1179 δοκῶν ἀποίσειν, αὐτὸς ἔνθεν ἦν: ὁ δὲ

1180 κάκ' εἰς μέγιστ' ἔσωσεν. εἰ γὰρ οὗτος εἶ

1181 ὃν φησιν οὗτος, ἴσθι δύσποτμος γεγώς.

Fig. 4: *Default view of the Greek drama.*

Οἰδίπους

1174 ὡς πρὸς τί χρείας;

ὡς πρὸς τί χρείας ; ὡς ἀναλώσαιμί νιν .

Θεράπων

1174 ὡς ἀναλώσαιμί νιν.

ὡς πρὸς τί χρείας ; ὡς ἀναλώσαιμί νιν .

Fig. 5: *An expanded line (*OT*, 1174).*

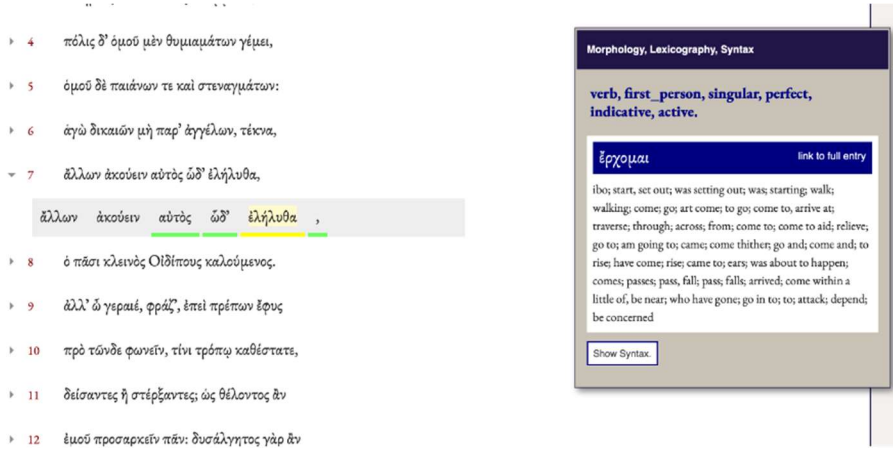

Fig. 6: *Hovering over a word.*

As a pedagogical tool, when reading collectively and remotely with students, some of them very new to Greek, Blackwell found it helpfully encouraging to invite the class to read a line in Greek, while pausing the cursor over each word, before calling on a student to translate. The students get a quick reminder of the vocabulary; when a student, translating, seems to hesitate, the teacher can simply point to the word that seems to be the problem, and the student (with no shame) has access on the screen to its morphology and lexicography. In a class where our goal was to understand Greek and think about the play, as opposed to testing day-by-day the students' ability to memorize, this allowed us to move quickly and have meaningful discussions.

In order to access the full *LSJ* entry, a reader must click on a word, as opposed to hovering over it. This will 'lock' the selection, allowing a click on the lexicon short definition. A second click will 'unlock' the selection, and the interface will go back to 'hover mode'.

12. *The User Interface and Syntax*

The basis for this whole project was to capitalize on an existing syntactic treebank, a comprehensive commentary on the play by an expert reader. This syntactic data is exposed to readers in two different ways.

In 'hover mode', the word currently under the cursor is highlighted with a yellow underline. Any words that are direct dependents on that word, according to the dependency formalism briefly introduced above, are underlined in Greek.

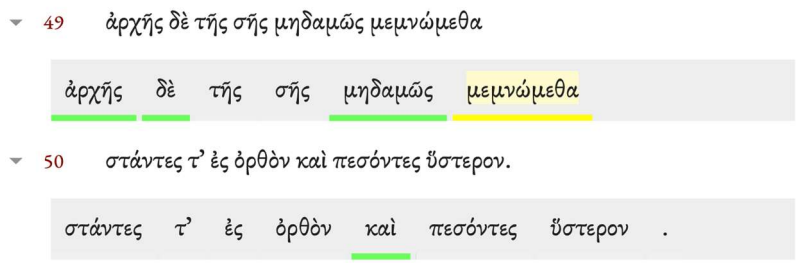

Fig. 7: *Hovering over the root of the sentence.*

In this example (fig. 7), from lines 49-50, μεμνώμεθα, 'let us remember', is the main verb, and thus the *root* of the syntax graph. It has three direct syntactic dependents: ἀρχῆς (the object), μηδαμῶς (a negative adverb), and καί (the immediate dependent that coordinates the two circumstantial participles στάντες and πεσόντες)[34].

[34] According to the dependency formalism adopted by the AGLDT, two or more coordinated elements depend on the (last, if the coordination joins more than two elements) coordinating conjunction. Coordination, with its 'horizontal' relations, is a notoriously difficult phenomenon to capture within the 'vertical' (*i.e.* head-to-dependent) dependency framework. See Osborne, *A Dependency Grammar* cit., pp. 297-320.

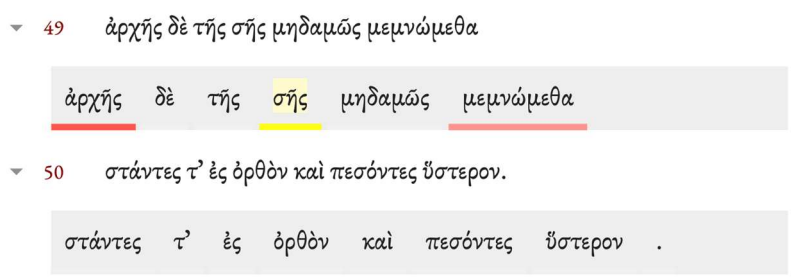

Fig. 8: *Seeing the chain of dependency.*

In the example of fig. 8, hovering over σῆς reveals the words on which it depends syntactically, underlined in red (ἀρχῆς). The immediate ancestor is darker red, with more remote generations showing increasingly fainter red underlines. This makes it clear that the path (back *up* the syntactic graph) is: σῆς … ἀρχῆς … μεμνώμεθα.

As part of the build process for the HTML pages, and based on the original enhanced dataset, the site includes provision for showing the syntax of any sentence in the "rooted, directional, acyclic graph" format familiar to anyone who has done dependency treebanking. This graph can be accessed from the **Morphology, Lexicography, Syntax** box in the upper-right. With any word highlighted, clicking 'Show Syntax' will reveal the whole graph of the sentence of which that word is a part.

The inline-indications of syntax were found to be helpful to readers and for line-by-line discussion of Sophocles' language. The tree-view proved especially useful for the more challenging passages. After students had become more familiar with the Greek, conversations based on syntax tended to move from «How does the sentence work?» to «Why did Mambrini make these choices? What else could he have done?».

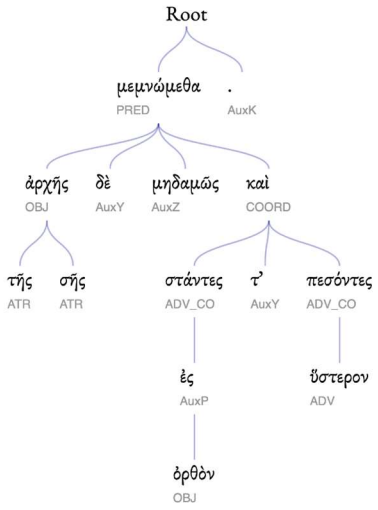

Fig. 9: *A syntax graph for* OT*, 49-50.*

13. *Final Pedagogical Notes*

> Cur ergo graecam etiam grammaticam oderam talia cantantem? nam et Homerus peritus texere tales fabellas, et dulcissime vanus est, et mihi tamen amarus erat puero. credo etiam graecis pueris Vergilius ita sit, cum eum sic discere coguntur ut ego illum videlicet difficultas, difficultas omnino ediscendae linguae peregrinae, quasi felle aspergebat omnes suavitates graecas fabulosarum narrationum. nulla enim verba illa noveram, et saevis terroribus ac poenis, ut nossem, instabatur mihi vehementer (Augustine, *Confessiones*, I, 14[35]).

[35] «Why then did I detest Greek literature when it told similar stories? For Homer too was skilled at weaving myths, and was just as delightfully vain; but

The authors recognize that the entire philosophy of this project runs directly counter to decades of Classical Language Pedagogy, which emphasizes memorization of vocabulary and morphological paradigms. A traditional 'reading class', such as marked both authors' education, whether in Europe (Mambrini) or the United States (Blackwell), involved a great deal of arduous reference to paper lexica and printed paradigms. Each act of 'looking up' took many moments. Blackwell has used this approach with undergraduates over the years, and some have had a good experience. But no class before 2020 CE has ever finished an entire Greek tragedy in one semester, reading all of the choral passages *and* having time to discuss *the contents* of the play. With this experimental edition, the most diverse Greek class in Blackwell's memory, under the worst possible circumstances, having been dispersed to their homes, far from the library or any immediate support, read every word of the play, discussed its words and syntax, challenged Mambrini's choices about syntax from time to time, and got to know the characters in the play and the universe they inhabited.

The last day of class in May, 2020 came, and the class was still about 100 lines short of the end of the play. The students insisted on additional meetings to finish reading.

14. *Looking Forward and Lessons*

This digital edition of the *OT* was an experiment, based on the happy accident of a previous class, under Blackwell's guidance, creating an enhanced dataset in a useful tabular format based on Mambrini's treebank.

when I was a boy I found him little to my taste. I suspect that Greek boys have the same reaction to Virgil, since they are made to learn him in the same way I learned Homer. Evidently there is difficulty, real difficulty, in learning a foreign language at all, as if it sprinkled all the sweet flavor of the Greek mythical stories with a foul taste. I knew none of the vocabulary, and I was severely intimidated by harsh threats of punishment to make me learn» (translation by C. Hammond, *Augustine. Confessions*, vol. I, Books 1-8, Harvard University Press, Boston 2014, p. 41).

It should be reproducible for any play, or any work of literature for that matter, for which we have treebank data. The main challenge, for generalizing production of editions like this, would be twofold. First, the creators will have to align lexical data from the treebanks, which are expressed in terms of (often ambiguous) lemmata, with explicit CITE2-URNs pointing to the online *LSJ*. And secondly, they will have to add speaker-attribution data.

The second task would be to refactor the bespoke code libraries that did implement the steps that we discuss above, namely:

- transform tabular data into CTS Editions and Exemplars
- generate the CITE Collections and Indices that aligned morphology and syntax
- built the HTML pages.

The developers of the CITE Architecture are currently working on code libraries for handling 'analyzed tokens', which are intended to make generic some of the specific work behind this *OT*.

This code library under development is based on the principle that 'it is always easier to aggregate than to de-aggregate.' Our work on the *Oedipus Tyrannos* has supported this principle. The hardest work was extracting and disambiguating the complex treebank XML. Once that data was in tabular format, building a complexly interactive website was a matter of weeks for one part-time enthusiast.

Finally, the world of philology needs better ways to re-use and re-purpose the incredibly valuable libraries of syntactic treebanks that scholars have created over the past two decades. We have profited from tools like Arethusa, the web application that allows users to upload Greek or Latin texts and produce treebank annotation using a neat visual interface[36]. But it remains the case that the default way to share treebank data is via screenshots (like the one in fig. 1), which is not ideal. On the contrary, we hope that our experience has shown once again the importance of reusing, attributing, citing and discussing linguistic annotation, both as a form of scholarly output, as a support for corpus-driven research, and as a pedagogical tool.

[36] See <https://www.perseids.org/tools/arethusa/app/#/>.

Teaching Oedipus Tyrannos

Abstract.
The article presents a digital scholarly edition of Sophocles' *Oedipus Tyrannos* that was used as support for an undergraduate seminar in Ancient Greek at Furman University. Our edition is a series of HTML pages, built programmatically, integrating syntactic, morphological, and lexical data. The edition is based on two of the most important acquisitions in the last decade of research in the Digital Classics: the CITE/CTS architecture and the treebanks of Ancient Greek. Our approach aims to overcome both the technical and linguistic "friction" that hinders the process of reading ancient texts in their original, unabridged format. While our work on the notion of a "zero-friction edition" predates the spread of COVID-19, all the problems that we were trying to tackle dramatically came into prominence with the outbreak of the pandemic during the Spring Semester of 2020. We discuss the outcomes and lessons-learned from this pedagogical experience.

Keywords.
Digital editions, pedagogy; treebanks, CITE architecture, Greek tragedy.

Christopher Blackwell
Furman University
christopher.blackwell@furman.edu

Francesco Mambrini
Università Cattolica del Sacro Cuore
francesco.mambrini@unicatt.it