Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 59–67
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

From Inscriptions to Lexicon and Back: A Platform for Editing and Linking the Languages of Ancient Italy

Valeria Quochi[1], Andrea Bellandi[1], Fahad Khan[1], Michele Mallia[1], Francesca Murano[2], Silvia Piccini[1], Luca Rigobianco[3], Alessandro Tommasi[4], Cesare Zavattari[1]
[1]Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche Italy, Pisa
[2]Dipartimento di Lettere e Filosofia, Università di Firenze
[3]Dipartimento di Studi Umanistici, Università di Venezia
[1]name.surname@ilc.cnr.it, [2]francesca.murano@unifi.it, [3]luca.rigobianco@unive.it
[4] Università di Pisa

Abstract

Available language technology is hardly applicable to scarcely attested ancient languages, yet their digital semantic representation, though challenging, is an asset for the purpose of sharing and preserving existing cultural knowledge. In the context of a project on the languages and cultures of ancient Italy, we took up this challenge. This paper thus describes the development of a user friendly web platform, EpiLexO, for the creation and editing of an integrated system of language resources for ancient fragmentary languages centered on the lexicon, in compliance with current digital humanities and Linked Open Data principles. EpiLexo allows for the editing of lexica with all relevant cross-references: for their linking to their testimonies, as well as to bibliographic information and other (external) resources and common vocabularies. The focus of the current implementation is on the languages of ancient Italy, in particular Oscan, Faliscan, Celtic and Venetic; however, the technological solutions are designed to be general enough to be potentially applicable to different contexts and scenarios.
Keywords: Digital Epigraphy, Restsprachen, Lexicon Editing and Linking

1. Introduction

Many languages spoken in antiquity have reached us through written testimonies that, in some cases, can be extremely limited both quantitatively and qualitatively. For these languages the denomination of *Restsprachen* 'languages of fragmentary attestation' is used, since their corpora can consist of a very small number of texts, even a few dozen, mostly typologically limited to the epigraphic form (inscriptions, stamps, coin legends). In terms of content, *Restsprachen* documentation is limited to the areas in which writing was selected by a given socio-cultural environment. The randomness of the findings amplifies the situation of fragmentation and precariousness of the knowledge we have of these linguistic systems, whose reconstruction is substantially partial, both in terms of grammar and lexicon, and limited in their sociolinguistic and diachronic complexity. It is often impossible to have a complete attestation of a declension or paradigm or to understand in depth the semantics of a form. This state of partiality has an impact, for example, on the lexicographic side, since lemmatization operations cannot take place appropriately, so it is necessary to resort to alternative forms of representation. Furthermore, the nature of the attestations makes an epigraphic approach to documentation indispensable.

Clearly, available language technology is hardly applicable without adjustments to this kind of languages because of both the high degree of uncertainty and data scarceness, which makes current machine learning and neural systems ineffective. Nevertheless, digital formalization and semantic representation of *Restsprachen* is an asset *per se* for the purpose of sharing and preserving existing knowledge. Setting up user-friendly digital tools that facilitate a full explicit encoding of available linguistic knowledge of these kind of languages according to up-do-date common models is certainly a challenge, but is, at the same time, important for bridging the digital gap and making the available knowledge and documentation widely accessible across disciplines.

This contribution takes up the challenge and describes the development of a user friendly web application, called EpiLexO, for the creation of lexica for fragmentary ancient languages with linking to the texts in which they are attested, as well as to bibliographic data and other (external) resources. The focus of the current implementation is on the languages of ancient Italy, as the platform comes to life within the project "Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models" (ItAnt hereafter)[1], which aims at investigating the cultures of ancient Italy on the basis of their relevant linguistic documentation bringing together methods and practices from traditional linguistics, philology, and digital technology. The technological solutions devised, however, are designed to be general enough to be potentially applicable to other contexts as well.

The paper is organized as follows. Section 2 sketches some background and related works; section 3 describes the platform design: the data encoding models and choices are described in §3.1 and §3.2, while the platform architecture is sketched in section 4. Section 5 describes the GUI by means of an exemplar use case.

[1]A project funded by the Italian Ministry of University and Research under the PRIN 2017 programme.

Finally, we conclude by sketching the scheduled future developments in section 6.

2. Background and Context

In the last decade, work on digital epigraphy was intense and, online platforms have flourished. It would be impossible in the limited space of a section to review all relevant experiences; therefore, we will briefly focus on the sources of inspiration for the present work.
The EAGLE project[2], part of Europeana, collects a massive amount of resources related to ancient world inscriptions, making them available for personal or research use. Its online platform allows for advanced searches over various databases such as the Epigraphic Database of Bari (EDB), the Epigraphic Database Heidelberg (EDH), and the Epigraphic Database Rome (EDR). It currently represents a reference within the digital epigraphy community, both in terms of quantity of materials made available and in terms of knowledge shared as open common vocabularies that, notwithstanding some limitations, are widely adopted.
Papiry.info[3] is a platform that contains a large collection of digital texts of Greek and Latin papyri and consists of two main components: a tool for searching and browsing the documents, and an editor which allows scholars to easily contribute to the collection by either creating new digital editions or proposing revisions of existing ones.
i.Sicily[4] offers a rich corpus of digital critical editions of inscriptions from ancient Sicily and an attractive web interface for the fruition of the digitized materials (Prag and Chartrand, 2019). The project has pushed the state of the art in digital epigraphy towards current language technology standards, such as the TEI EpiDoc XML format for digitizing inscriptions and partly towards the semantic web. Unlike most other similar initiatives, i.Sicily does not focus on texts only, rather these are enriched with bibliographic references and other metadata, such as person and geographic names from the Pleiades and Geonames vocabularies, and Trismegistos IDs.
Last but not least, the recent Cretan Institutional Inscriptions (CII) project delivers an EpiDoc XML database of inscriptions and offers an online search and consultation interface based on the EFES front-end service (Bodard and Yordanova, 2020). In addition to text encoding and the adoption of Linked Open Data (LOD) common vocabularies, the database includes cross-linked bibliography and various indices to allow for quick search on the contents (Vagionakis, 2021).
In most of the reviewed projects, language, i.e. the lexicon, is the great absentee. In ItAnt we therefore chose to focus on complementing the digital epigraphy landscape with tools for creating Semantic Web compliant ancient lexica and integrate them with texts and other online datasets.
The publication of language resources for ancient languages on the Semantic Web is still at a fairly early stage. One pioneering work on this topic is certainly the Linking Latin (LiLa) project[5], which created a knowledge base of linguistic resources for the Latin language, and publishes numerous such resources. One of the most innovative aspects of this knowledge base is that the different resources (lexica and corpora) are all linked together via lemmas (the core of the LiLa Lemma Bank). This, together with the use of standardized models for representing different resources (such as OntoLex-Lemon and its extensions for representing lexicons), ensures that the entirety of the knowledge base is interoperable both internally and externally.
One issue that arises frequently in the modeling of resources for ancient or historic languages is the necessity of representing etymological derivation. Although a consensus has not yet been reached within the linked data community, strategies for dealing with this have been proposed. Khan (2018) proposes an OntoLex-Lemon compatible vocabulary, *lemonEty*, for representing etymologies as hypothetical word histories; albeit not official, the *lemonEty* extension is the solutions adopted in LiLa (Mambrini and Passarotti, 2020).
Another issue very pertinent to humanities use cases is the linking together of lexicons with corpora, usually in order to represent the attestation of a lexical element in a corpus. From the lexical point of view this is the topic of a new set of specifications (currently in progress) designed to extend the OntoLex-Lemon guidelines with classes and properties for, among other things, representing such links; these are the Frequency Attestations and Corpus (FrAC) specifications (Chiarcos et al., 2020) [6].

3. The Platform

EpiLexO is a platform dedicated to the creation and editing of lexical resources for ancient fragmentary languages integrated, i.e. linked, to their 'testimonies' (i.e. transcriptions of epigraphic texts), to related bibliography, to contextual metadata, and to other relevant independent (LOD) resources, such as the LiLa Knowledge Base (Mambrini et al., 2020) and common vocabularies. Its implementation is based upon current standards in software design and relies on previous experiences within the Digital Humanities (DH) and Language Technology (LT) communities (cfr. §4 below). It is realized as a SOA system with strong frontend-backend separation of concerns in such a way that makes most services potentially reusable in different contexts. The web application is conceived to allow for a dual mode: (1) an 'edit mode' which allows for the editing of lexical data and its linking to the various external resources; and (2) a 'view mode', which will

[2] https://www.eagle-network.eu/
[3] https://papyri.info//
[4] http://sicily.classics.ox.ac.uk/
[5] https://lila-erc.eu/
[6] https://www.w3.org/community/ontolex/

allow users to search and study the digitized materials by cross querying on the different datasets. It shall also provide export functionalities to download data in LOD compliant formats.

In this contribution we describe its first implementation for dealing with the highly fragmentary attested languages of ancient Italy. Although still work-in-progress, the α version of the editing mode is complete. The intended users are historical linguists, expert in one of more ancient language(s).

In this platform, the lexicon is pivotal, as the focus of the whole system is language; text is here seen as instrumental for the construction and enrichment of the lexicon. Hence, the platform does not include text editing functionalities, rather it assumes the existence of a suitable corpus to be ingested. In practice, within the ItAnt project texts are encoded independently as described here below[7].

3.1. TEI EpiDoc

Within the ItAnt project, texts are encoded independently of the platform according to the TEI/EpiDoc guidelines[8], the de-facto standard for digital epigraphic projects, in order to create a digital edition of the ItAnt corpus by providing information concerning text both as a linguistic and a material object in a semantic format. Each inscription is described in its archeological, epigraphic and linguistic data; bibliographical references, commentaries and facsimiles are also provided. Concerning the identifiers section, we have chosen to include references to Trismegistos[9] and to the most important inscription collections, e.g. (Rix, 2002). The description of the support is enriched by reference to the Getty Vocabularies[10] in relation to the archeological object bearing text information (object type, material, execution tecniques), and to the EAGLE vocabularies[11] in relation to the inscription type. Finally, Pleiades[12] and Geonames[13] thesauri are used for respectively ancient and modern place-names. As an important innovation, every token in the inscriptions is marked as <w> and identified by the <xml:id> tag, to improve linkability. An example of the ItAnt text encoding can be seen if Fig.3.1 below, which contains a fragment of a Samnite instription.

Compared to texts concerning classical languages, ItAnt corpus requires a specific markup regarding some elements, for the presence of non-classical epigraphic uses which would otherwise remain inaccurately described[14]. For instance, an accurate information about the token separation is required especially for systems like the Venetic alphabets (which show an inter-syllabic separation). In such cases, we chose specific values for the @type of the <tei:rs> tag. Furthermore, important elements and specific linguistic problems are not always sufficiently taken into account by the EpiDoc guidelines: for example, EpiDoc does not offer the possibility to distinguish the identification of the writing system from the description of the language. This is an important conceptual distinction from a linguistic point of view, and is also relevant in the study of texts, since the documentation of a language can also be written using different scripts. For instance, the Oscan corpus is written using a national Etruscan-based and a modified Greek alphabets, and lately the Latin one. According to the EpiDoc recommendations, the @ident attribute of the <language> element describes the scripts as connected to a language. To discern these two aspects, we chose to describe the writing system within the <tei:scriptDesc> tag, using the @ref attribute to link the concepts of the vocabulary of ancient Italy scripts the ItAnt project is creating. The <language> tag is used only for the representation of the languages[15].

3.2. The EpiLexO Lexical Model

The modeling of *Restsprachen* constitutes the springboard to tackle a number of lexicographic issues raised by the adoption of models that have been mainly designed for widely attested living languages. Differently from other lexical resources, notably from the Lila knowledge base, the core of the Epilexo model, based on Ontolex-Lemon, is constituted by word forms. The fragmentary attestation of Italic languages, as mentioned in §1, often makes it impossible to identify the lemma, i.e. the conventional form chosen to represent the lexical entry and used for normalization purposes. Word forms in EpiLexo correspond to reconstructed orthographic representations and function as the hook for the linking to the textual elements, i.e. to the transcriptions of epigraphy texts, to bibliographic references and to external databases. Although word forms play a central role in the ItAnt lexicon, our knowledge of the morphology is often very limited and our analysis can be compromised by the fact that many of these forms are uncertain, as documented by inscriptions severely damaged by time. Thus, for example, in the inscription ItAnt_Osc_3 the form *legú* is expanded by some editors as *legú(m)*, which is to be interpreted as the

[7] This choice was also based on the consideration that work on TEI XML editors is presently quite advanced and that there might be opportunities in the future to integrate with them, rather then compete, see for instance Janssen (2016), Del Grosso and Nahli (2014) and Del Grosso (2015). In fact, an experiment within ItAnt is ongoing for encoding texts in TEI Epidoc with a Domain Specific Language based on the EUPORIA system (Boschetti and Del Grosso, 2018).

[8] https://epidoc.stoa.org/gl/latest/intro-intro.html//

[9] https://www.trismegistos.org//

[10] https://www.getty.edu/research/tools/vocabularies/aat///

[11] https://www.eagle-network.eu/resources/vocabularies/typeins//

[12] https://pleiades.stoa.org//

[13] http://www.geonames.org//

[14] Similar problems have also been addressed by the ILA project for the encoding of archaic Latin inscriptions (Sarullo, 2016)

[15] A paper focusing on the EpiDoc encoding of inscriptions by the ItAnt project is being prepared.

Figure 1: A fragment of the ItAnt_Oscan_3 edition of the Sa 2 inscription

```
<tei:div type="edition" subtype="interpretative" xml:space="preserve">
  <tei:div type="textpart" n="face_a" style="text-direction:r-to-l" rend="ductus:sinistrorse">
    <tei:ab>
      ....
        <tei:name type="patronymic" xml:id="Osc_3_l_1_w_6" ref="#p2"><tei:expan>
           <tei:abbr><tei:unclear>st</tei:unclear></tei:abbr><tei:ex>aatieís</tei:ex>
           </tei:expan>
        </tei:name>
        <tei:w xml:id="Osc_3_l_1_w_7">legú</tei:w>
        <tei:pc unit="word">.</tei:pc>
        <tei:w xml:id="Osc_3_l_1_w_8">tangi<tei:unclear>n</tei:unclear>úd</tei:w>
      <tei:lb n="2" xml:id="Osc_3_l_2"/>
        <tei:w xml:id="Osc_3_l_2_w_1">aam<tei:unclear>a</tei:unclear>n<tei:expan>
        <tei:ex>a</tei:ex></tei:expan>fed</tei:w>
        <tei:pc unit="word">.</tei:pc>
        <tei:w xml:id="Osc_3_l_2_w_2">e<tei:unclear>s</tei:unclear>í<tei:supplied
        reason="lost" evidence="previouseditor">dum</tei:supplied></tei:w>
        <tei:pc unit="word">.</tei:pc>
        <tei:w xml:id="Osc_3_l_2_w_3"><tei:supplied reason="lost"
        evidence="previouseditor">prúfat</tei:supplied><tei:unclear>e</tei:unclear>d</tei:w>
        <tei:pc unit="word">.</tei:pc>
        <tei:w xml:id="Osc_3_l_2_w_4"><tei:unclear>ú</tei:unclear>psed</tei:w>
        <tei:pc unit="word">.</tei:pc>
        ....
    </tei:ab>
  </tei:div>
</tei:div>
```

genitive plural of *leg-* 'law', while others expanded it as *legú(túm)*, which is to be interpreted as the genitive plural of a noun denoting a public institution. In order for the linguistic information to be reliable, it is therefore crucial to link lexical information with corpus evidence. To this end, we adopted some classes and properties, which are being developed as part of the FrAC extension to Ontolex, as mentioned in section 2 above. More specifically, each form of a lexical entry is associated to its exact occurrence(s) in the ItAnt digitized inscription(s) (if the epigraphy has been digitized and transcribed), or generically to the inscriptions it is attested in. The Attestation can then be further enriched with additional information, e.g. about whether the form reading is conjectural.

Etymological information is modeled with the *lemonEty* extension, mentioned in §2. For each lexical entry it is possible to specify either or both the Proto-Italic and Proto-Indo-European reconstructed forms (encoded as instances of the class *Etymon*, i.e. Lexical Entries with a special status) as well as the cognate words attested in sister languages (instances of the class *Cognate*). In order to specify the type of etymological derivation process, and because a common owl vocabulary for etymological knowledge is missing, we borrow the values of *etyLinkType* from the Lexical Mark-up Framework (as described in the normative Annex B of LMF Part 3). Specifically, *inheritance* and *borrowing* make it possible to define if we are dealing respectively with a direct hereditary relation from an ancestor or rather with a word borrowed from another genetically (un-)related language. In accordance with the Linked Data principles, the Latin cognates as well as the etymological roots may also be linked to e.g. the LiLa knowledge base, either via the *seeAlso* and *sameAs* relations, or directly. Although the system is set up for modeling different etymological hypotheses, for the time being it is up to lexicographers to choose the reconstruction that they deem most reliable. Some non standard properties are introduced to formally describe specific features, such as the data properties *stemType* to indicate which thematic class the lexical entry belongs to (e.g. *ā-* stems which are stems ending in *-$\bar{a}$ < PIE *-*eh2* belonging to a specific declension type),[16] and *Uncertain* for expressing uncertainty at the level of morphology, sense and etymology. The class *Bibliography* – along with a set of properties – makes it possible to specify author, title, data, pages of the bibliographic references and to include the link to the Zotero database (see infra). Currently, it is a system-internal data structure not yet mapped to any common vocabulary. It will be rethought in the light of the new IFLA Library Reference Model (IFLA-LRM) (IFLA Functional Requirements for Bib-

[16] A similar non standard choice is done in LiLa, which defines a specific ontology for describing morphological properties of word formation, while waiting for the Ontolex morphology extension to take a definitive shape. In ItAnt it was decided that knowledge of the language systems is not yet mature for a proper modeling of this aspect, and temporarily to encode such information as a data property, although in principle its values belong to a closed class

liographic Records (FRBR) Review Group: Riva, P. et al., 2018). Examples from the Oscan lexicon will be given in §5.

4. EpiLexO: A Sketch of the Architecture

EpiLexO follows a REST architectural style, where the implementation of the client and the implementation of the servers are done independently. The server side is composed of two main back-ends, namely the LexO-server and the CASH-server, which manage lexica and text documents respectively. They expose APIs based on the HTTP protocol and exchange data in JSON format. The services conform to OpenAPI, a specification for machine-readable interface files to describe, produce, consume and display REST services.

LexO-server[17] stands for **Lex**icon and **O**ntology-server and has evolved from the experience of LexO-lite (Bellandi, 2021), a full stack tool for editing OntoLex-Lemon resources. The LexO-server allows for managing both linguistic and conceptual dimensions, and for a correct linking between each other, according to either a semasiological or an onomasiolgocal approach. Concerning the linguistic part, LexO-server heavily relies on the *OntoLex-Lemon* model, while the conceptual one is based on Simple Knowledge Organization System (SKOS). LexO-server is written in Java and uses a semantic repository called GraphDB.

CASH-server stands for **C**orpus, **A**nnotation, and **S**earc**H**-server[18]. It exposes a set of services for, i) managing a corpus of text documents and organize it like a file system; ii) linking corpus and lexicon, i.e., creating annotations that represent the linking of lexical elements to text portions (defined by spans of characters), with associated metadata (e.g. author, confidence, bibliography, etc.); iii) making multilevel searches involving lexicon, texts, links, and metadata. Annotations can refer to any span of characters and thus equally relate to words, subwords and multiwords. The same span can be annotated multiple times, thus allowing for the piling up of an arbitrary number of annotations, which may correspond to different descriptive layers, or concurrent alternatives. CASH is devised to be general and modular, in particular concerning the import functionalities,[19] conceived as plug-in ingestion module that may manage different file formats. At present the system supports the importing of EpiDoc-XML[20].

The EpiLexO platform also relies on external REST APIs, e.g. on Zotero[21] for associating bibliographical references to lexical items and attestations; and on KeyCloack[22] for user management, authentication and authorization.

EpiLexO GUI. The services exposed by the servers are invoked by an interface developed in Angular[23] and designed as a single-page web application made up by several components. Each component offers different functionalities for creating lexical items and (inter)linking them with other internal or external data (i.e. lexicon items, corpus texts, bibliography, vocabularies, LD resources)[24]. All interface components communicate with each other through the use of services based on RxJs technology[25], a library integrated in Angular for event-based programming and asynchronous call management.

The platform GUI, shown in Fig.2, is divided into three main vertical sections, dedicated to a set of different kinds of activities.

The left column (a) is subdivided into three panels and shows the navigation trees for the main resources: corpus, lexicon, ontology[26], each one with its peculiar structure and functionalities.

The central part (b) is the main working area devoted to the editing tasks; the lower part contains the lexicon editor, the upper part the text linker. The lexicon editor is pivotal to the whole platform, modular and contextually adaptive, i.e. it shows editing sections on the basis of the item selected in the lexical entry tree, and editing is dynamic, that is changes in the values are directly recorded and registered in the back-end. As Epilexo presently makes use only of a subset of the Ontolex model (cfr. §3.2 above), it allows for the encoding of information for lexical entries, forms, senses and etymologies.

The right column (c) contains several panels, dedicated to various kinds of "accessory information": metadata about the (edition of the) inscriptions, free textual notes, links to external resources, bibliographic references, attestations. The content of these panels is also contextual. i.e. dynamically dependent on the items selected in the left or central column.

In the following section the platform will be described into some details by means of examples based on the first bulk of the Oscan lexicon[27].

[17] https://github.com/andreabellandi/LexO-backend

[18] https://github.com/valeq/backendLexO-textAnnotations

[19] A paper focusing on the system of back-ends is in preparation, which will also discuss their potential for application in other DH scenarios.

[20] In fact, as there are several possible equally valid Epidoc-XML dialects, and given the peculiarities of the ItAnt variant, at the moment the system ingests the ItAnt Epidoc format. However, the XML importer is designed to be customizable relying on xpath syntax, and adaptation to different XML formats shall be possible. This has still to be tested.

[21] https://www.zotero.org/

[22] https://www.keycloak.org/

[23] https://angular.io/

[24] https://github.com/MicheleMallia/LexO-angular

[25] https://rxjs.dev/

[26] The services offered by the LexO-server for importing an external ontology and link its elements to lexical items, are currently not exploited in ItAnt.

[27] Encoded for ItAnt by Dott. Edoardo Middei

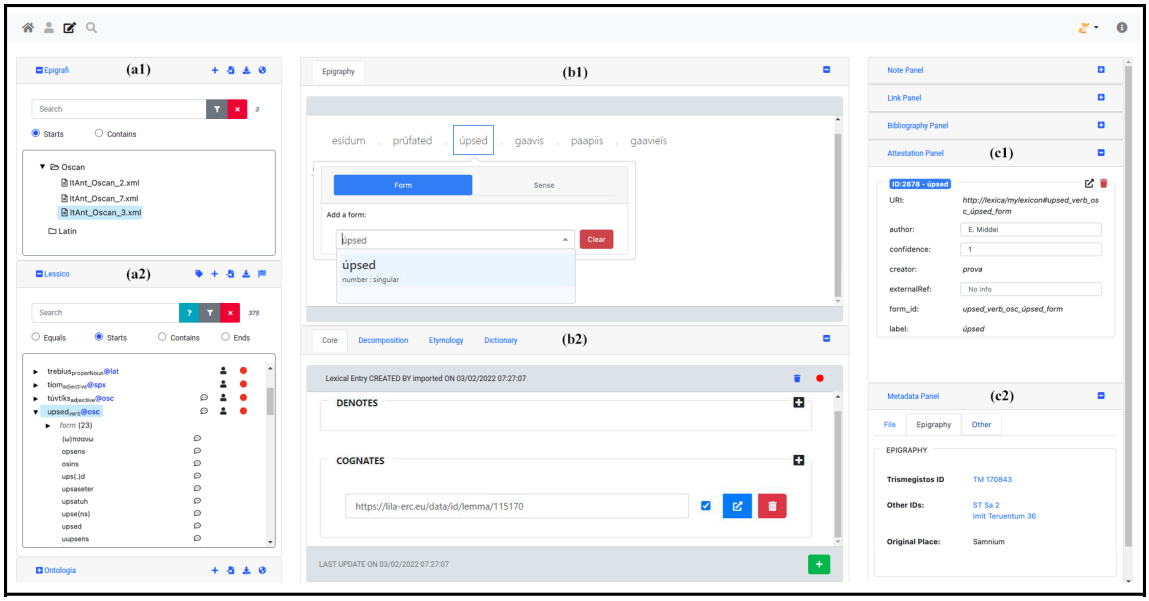

Figure 2: The EpiLexo GUI

5. Editing the Lexicon and Linking the Inscriptions

EpiLexo accommodates two possible usage workflows: 1) creation of a lexicon from scratch on the basis of (a corpus of) epigraphic texts; and 2) linking of an existing lexicon to (a corpus of) epigraphic texts in which its languages are attested. In the first scenario, a scholar imports (an EpiDoc XML corpus of) epigraphic texts into the platform and starts creating and encoding the lexical items attested in the corpus, linking them to the exact textual loci, to relevant bibliography and possibly to relevant external resources. In the second scenario the scholar starts from an existing lexicon for the language(s) of interest, imports a corpus and encodes the links to the various relevant internal and external data. Possibly (s)he can further edit and enrich the lexicon by adding e.g. new entries and forms[28].

The platform permits to perform all the required actions from a single page in a seemingly smooth way.

From within the corpus panel in the left column, the user uploads a corpus file that documents one or more lexical entries, for instance the ItAnt_Oscan_3.xml (excerpted in §3.1 above), which represents a critical edition of a Samnite inscription. The corpus panel –(a1) in Fig.2– is organized like an OS file system and allows for typical CRUD operations, based on the CASH-server APIs. Thus, the user can create folders, move files, import other Epidoc documents, add metadata tags to both files and folders. The importer automatically extracts from the EpiDoc file all metadata related to the inscription and its edition, and the platform dynamically displays them in the metadata panel in the right column, as in Fig.2 (c2). In our example, for instance one can easily check e.g. the inscription provenance (Samnium) and the other identifiers by which the inscription is also known as (i.e. TM 170843, Sa 2, Teruentum 36)[29].

The text contained in the inscription is shown in the central upper panel (Fig.2 (b1)), the Linker, which allows for linking text portions to items in the lexicon by invoking the services of the CASH-server. Because the ItAnt Epidoc corpus encodes word segmentation (cfr. §3.1), the Linker makes use of this information and displays the text into visual segments. Linking is done by selecting an entire token, a subtoken (e.g. a prefix), or a list of tokens (for linking to multiwords), and then searching for and selecting the desired form from the lexicon within a dedicated pop-up window, as shown in the figure for *úpsed*.

The act of establishing a link between a text portion and a lexical form practically corresponds to creating an Attestation for the given form, according to the model described in §3.2 above. Attestations are displayed in the dedicated panel on the right column, (c1) in the figure, from where they can be further enriched according to the model. For instance, in the case of *úpsed* we may want to set the confidence to 1 to assert certainty, attribute authorship to a different scholar, add a bibliographic reference, e.g. to Untermann (2000) where the specific attestation is discussed (see Fig.5 in the Appendix for an example of the Zotero plug-in for adding blibliographic references to the lexicon).

[28] In ItAnt, this second scenario is the actual case for the Oscan language, for which the lexicon encoding started with an ad-hoc adaptation of LexO-lite (Bellandi, 2019).

[29] Notice that the identifiers are displayed as hyperlinks pointing to the actual external resources, i.e. to the Trismegistos record and to the bibliographic records of the secondary sources in the ItAnt Zotero library[30]

Before working on the linking, (s)he might want to first check how the lexical entry for *upsed* is encoded in the lexicon. S(he) would then use the filter in the lexicon navigation tree in the left column dedicated to managing and browsing the lexicon content (cfr. Fig.2.(a2)). This panel is organized according to the key ItAnt lexical classes: Lexical Entry, Form, Senseand Etymology (cfr. §3.2), which dynamically correspond to dedicated editing views in the central part of the interface. From this panel the user can also perform some high-level lexicon editing actions, such as adding new languages and lexical entries.

Because of the theoretical and practical difficulties of lemmatization discussed in §1 and §3.2 above, forms are richly described and represent the key elements in the lexicon, acting also as the interface with the texts. In the current version, forms are all listed and grouped under a Lexical Entry, as it can be seen in Fig.3. Information about whether the lexical entry is an etymon (i.e. an etymological root), about its stem type and about its cognates is also encoded at lexical entry level, which is to be considered as a mere container for encoding those features shared by all forms (such as language and part of speech).

Cognates are encoded by linking either internally to another entry of a different language or externally to another linked data compliant lexicon. In Fig.2 (b2) we see the Latin cognate of *upsed* represented by the URI of the corresponding lemma entry, *opus*, in the LiLa knowledge base[31].

Etymological information is attached to a Lexical Entry and applies to all of its forms. Etymology has a dedicated structure which, in addition to the etymon, allows for the specification of the type of derivation and the author[32]. Although the underlying model is capable of representing derivation chains, this possibility is deliberately blocked in the current interface on theoretical basis that need further reflection and confrontations. In Fig.3 we see an example of this: here the source and target of the etymological link are default values set by the system, as they always correspond to the etymological PIE or PIT root and the current lexical entry respectively. In principle, these fields can be made editable to permit the encoding of derivation chains.

Similarly to Cognates, the PIE etymon here can link either externally to the corresponding etymon in the LiLa Etymological Dictionary, or internally to the **h3ep-* Etymon entry, as exemplified in Fig.4. In the latter case, linking to the LiLa equivalent can be encoded at lexical entry level, in the Link panel on the right column by means of a `owl:sameAs` relation, as in Fig.4.

Bibliographic references to relevant literature can be added to lexical entries, forms, senses, etymologies, as well as to Attestations, via a Zotero plug-in (see Fig.5 in the Appendix) and enriched with additional information in the Bibliography panel, in the right column.

Finally, free textual notes for describing any additional unstructured, information can be added in the Note panel on the right column to every element of the lexicon; the same applies to links to relevant external resources, which can be encoded in the Link panel as `rdf:seeAlso` or `owl:sameAs` relations for any lexical element, as in the Etymology example above.

6. Conclusion and Future Works

In this paper we have presented a newly developed editing platform for the creation of interlinked linguistic datasets for ancient fragmentary languages. While the front-end is in part tailored on the specific requirements of the project it is born to serve, the whole architecture is modular and general enough to serve other needs as well. As mentioned above, the platform is not yet complete. The 'edit mode' is about to go in production and user feedback will prove precious for bug fixing and improvements. In the immediate future efforts will be devoted to the construction of the 'view mode', which should allow multi-layer, cross-dataset queries, as well as effective presentation of the contents and search results. In this respect, plans for CASH are to experimentally support a query language based on CQL that will permit to perform complex queries mixing text content with both metadata and annotations, such as: "find all inscriptions in language *L* (metadata), containing the word *W* (content) as an attestation of the form *F* found in the lexicon (annotation), followed by a person name (content+annotation)".

Another fundamental aspect that needs to be dealt with soon is export functionalities. As one of the objectives is to produce and publish a LOD version of the results, the platform shall allow for the exporting of the data in LOD compliant formats. While the lexicon will require only minor adjustments to be fully compliant to Ontolex, we still need to make decisions on the representation of texts, bibliography, and bibliographic references or citations. For the latter, good candidates are the FRBR-aligned Bibliographic Ontology (FaBiO) or the Citation Typing Ontology (CiTO) (Peroni and Shotton, 2012), while for the bibliography the IFLA-LRM mentioned in §4 will have to be assessed. As far as texts are concerned, internal discussion is still open; one safe but sub-optimal solution might be to follow the example of the LiLa knowledge base that provides a Powla rdf representation of texts as lists of tokens. However, this is a hot research topic in the humanities

[31] The choice of whether to link externally or internally to one of the lexicon entries is left to the scholar, and mostly depends on the availability of LOD compliant lexical resource for the language(s) of interest.

[32] Given that a Lexical Entry is allowed to have many Etymology items, the possibility to state the author might be used to encode alternative hypothesis, and goes in the direction suggested in Mambrini and Passarotti (2020) of treating etymologies as scientific propositions and model them also according to CIDOC CRM-tex (Felicetti and Murano, 2021). The current implementation however leaves this under-specified and conforms to the project requirement to encode only the editors' scientific claim(s).

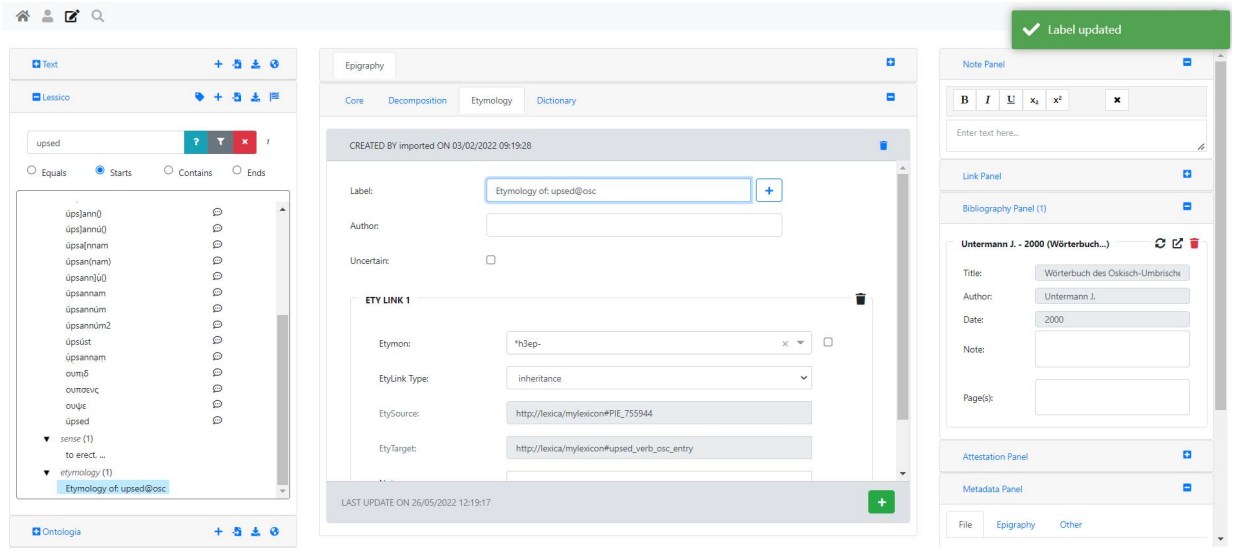

Figure 3: Etymology of *upsed*

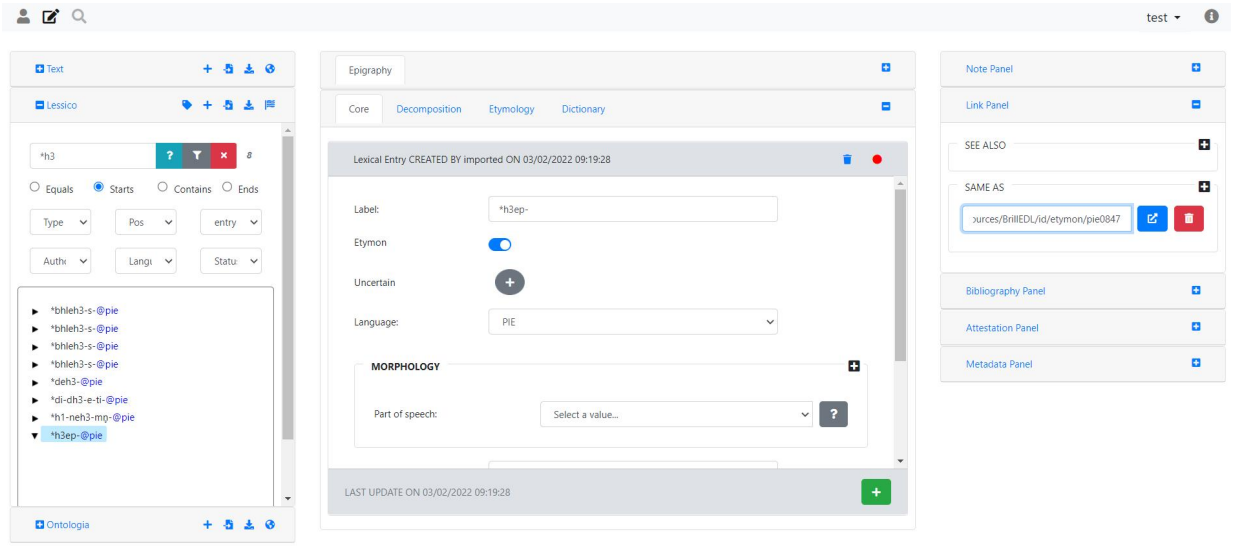

Figure 4: Linking Etymons to the LiLa Etymological Dictionary

and other options still have to be taken into account. The software is open source and, once complete, the full package will also be delivered on a docker image that can be quickly installed on any server. Finally, all results – corpus and lexical data as well the software – will be deposited in the ILC4CLARIN repository and integrated as a service into CLARIN-IT, which will guarantee long term preservation of the digital project outputs and the sustainability of the platform. To this end, work is in progress towards the integration with the CLARIN AAI and SSO services, via the Keycloak backend.

7. Acknowledgments

This work is financially supported by the Italian Ministry of University and Research under the PRIN 2017 programme and carried out within the "Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models" project (PRIN 2017XJLE8J).
It also has the support of the CLARIN-IT infrastructure.

8. Bibliographical References

Bellandi, A. (2021). LexO: an open-source system for managing ontolex-lemon resources. *Language Resources and Evaluation*, 55.4:1093–1126.

Bodard, G. and Yordanova, P. (2020). Publication, testing and visualization with EFES: A tool for all

stages of the EpiDoc XML editing process. *Studia Universitatis Babeș-Bolyai Digitalia*, 65(1):17–35.

Boschetti, F. and Del Grosso, A. M. (2018). Euporia: Piattaforma digitale per l'annotazione tramite Domain Specific Languages di testi multilingui disposti in parallelo.

Chiarcos, C., Ionov, M., de Does, J., Depuydt, K., Khan, F., Stolk, S., Declerck, T., and McCrae, J. P. (2020). Modelling frequency and attestations for ontolex-lemon. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 1–9.

Del Grosso, A. M. and Nahli, O. (2014). Towards a flexible open-source software library for multi-layered scholarly textual studies: An arabic case study dealing with semi-automatic language processing. In *Proceedings of the Third IEEE International Colloquium in Information Science and Technology (CIST)*, pages 285–290.

Del Grosso, A. M. (2015). *Designing a Library of Components for Textual Scholarship*. Ph.d. thesis in computer engineering, unpublished phd thesis, University of Pisa.

Felicetti, A. and Murano, F. (2021). Semantic modeling of textual entities: The CRMtex model and the ontological description of ancient texts. *Umanistica Digitale*, (11):163–175, Jan.

IFLA Functional Requirements for Bibliographic Records (FRBR) Review Group: Riva, P., Le Boeuf, P., and Zumer, M. (2018). IFLA library reference model: A conceptual model for bibliographic information. https://repository.ifla.org/handle/123456789/40.

Janssen, M. (2016). Teitok: Text-faithful annotated corpora. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Khan, A. F. (2018). Towards the representation of etymological data on the semantic web. *Information*, 9(12):304.

Mambrini, F. and Passarotti, M. (2020). Representing etymology in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 20–28, Marseille, France, May. European Language Resources Association.

Mambrini, F., Cecchini, F. M., Franzini, G., Litta, E., Passarotti, M. C., and Ruffolo, P. (2020). LiLa: Linking Latin. risorse linguistiche per il latino nel semantic web. *Umanistica Digitale*, 4(8), Jan.

Peroni, S. and Shotton, D. (2012). Ontology paper: Fabio and cito: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17:33–43, dec.

Prag, J. R. W. and Chartrand, J. (2019). I. Sicily: Building a digital corpus of the inscriptions of ancient sicily. In Annamaria De Santis et al., editors, *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*, pages 240–252. De Gruyter Open Poland.

Rix, H. (2002). *Sabellische Texte. Die Texte des Oskischen, Umbrischen und Südpikenischen*. C. Winter, Heidelberg.

Sarullo, G. (2016). The encoding challenge of the ILA project. In Antonio Felle et al., editors, *Off the Beaten Track. Epigraphy at the Borders. Proceedings of the VI EAGLE International Event (Bari, 24th-25th September 2015)*, Oxford. Archaeopress.

9. Language Resource References

Andrea Bellandi. (2019). *LexO - Lexicographic Editor for Ontolex-lemon Resources*. Istituto di Linguistica Computazionale "A.Zampolli", Consiglio Nazionale delle Ricerche, Italy, Pisa, PID http://hdl.handle.net/20.500.11752/ILC-95.

Irene Vagionakis. (2021). *Cretan Institutional Inscriptions Dataset*. Venice Centre for Digital and Public Humanities (VePDH), PID http://hdl.handle.net/20.500.11752/OPEN-548.

10. Appendix: the Zotero Plug-in

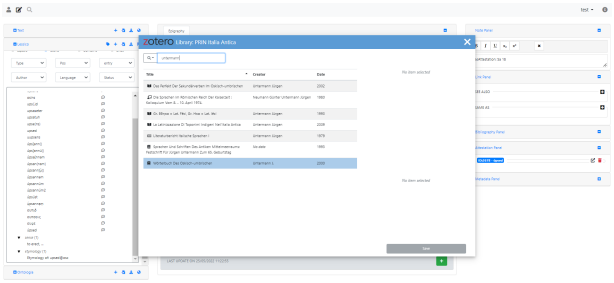

Figure 5: An example of the Zotero plug-in: searching the relevant bibliographic entry for Untermann 2000, to be linked to *upsed*.