Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2856–2864
Marseille, 20-25 June 2022

Placing multi-modal, and multi-lingual Data in the Humanities Domain on the Map: The Mythotopia Geotagged Corpus

Voula Giouli, Anna Vacalopoulou, Nikolaos Sidiropoulos, Christina Flouda, Athanasios Doupas, Giorgos Giannopoulos, Nikos Bikakis, Vassilis Kaffes, Gregory Stainhaouer
ATHENA Research and Innovation Centre
Artemidos 6 & Epidavrou, Maroussi, Greece
{voula, avacalop, nsidir, cflouda, adoupas, giann, bikakis, vkaffes, stein}@athenarc.gr

Abstract

The paper gives an account of an infrastructure that will be integrated into a platform aimed at providing a multi-faceted experience to visitors of Northern Greece using mythology as a starting point. This infrastructure comprises a multi-lingual and multi-modal corpus (i.e., a corpus of textual data supplemented with images and video) that belongs to the humanities domain along with a dedicated database (content management system) with advanced indexing, linking, and search functionalities. We will present the corpus itself focusing on the content, the methodology adopted for its development, and the steps taken towards rendering it accessible via the database in a way that also facilitates useful visualizations. In this context, we tried to address three main challenges: (a) to add a novel annotation layer, namely geotagging, (b) to ensure the long-term maintenance of and accessibility to the highly heterogeneous primary data – even after the life cycle of the current project – by adopting a metadata schema that is compatible to existing standards; and (c) to render the corpus a useful resource to scholarly research in the digital humanities by adding a minimum set of linguistic annotations.

Keywords: cultural heritage and humanities corpus, cross-media indexing, geo-tagging benchmark.

1. Introduction

The quest for digital resources in the humanities domain brings about the notion of representation and organization of the data at hand that will allow not only for the efficient indexing and retrieval of the digital content, but also for providing meaningful visualizations. Lately, adding accurate geographical information that is relevant to the content of texts, images, or videos has become of paramount importance not only to applications that seek to address the so-called *placing* task, but also to tools aimed at visualizing the spatial features of texts. When performed manually, geotagging is a costly and time-consuming endeavour. To overcome this obstacle, tools for geotagging are required, that is, tools to automatically disambiguate names of localities, identify locations and geotag texts, images, audio, or video. In this regard, there is a need for hand-crafted quality datasets that are applicable for training and evaluating purposes. The work presented here seeks to address this issue by creating a corpus resource that spans across various modalities, languages, and domains and its multi-layer processing by adding metadata – including spatial information.

The paper is outlined as follows: in section (2) we present the rationale behind corpus creation and the general outline of the project; previous work in developing corpora in the humanities domain and in developing geotagged datasets is summarised in (3). Section (4) gives an overview of the corpus, whereas the methodology adopted for the selection, development, and processing of the corpus along with the content management system developed is described in section (5). Section (6) gives an account of the geotagging procedure followed. The utility of the underlying infrastructure along with the geo-referenced material towards creating itineraries for tourists is showcased in (7), whereas, finally, our conclusions and plans for future research are presented in section (8).

2. Rationale and scope

The Mythotopia corpus is currently a work in progress, created in the wider context of developing an online platform offering a multifaceted view of Eastern Macedonia and Thrace in Northern Greece with a wealth of information from several points of view including mythology, history, architecture, natural environment, culture, society, folklore, recreation, gastronomy, travel and tourism, leisure, and more (Vacalopoulou et al., 2021). The problem of “multimodal location estimation” lies within the heart of the overall project. The platform uses different types of data to facilitate searches and retrieve information according to several separate or combined criteria, also offering the option of defining personalized itineraries in the area based on these criteria. Consequently, Points of Interest (POIs), i.e., geospatial entities that are characterised by at least a name and a set of coordinates and describe a place of some utility or interest to end users, are one of the core resources handled in the Mythotopia corpus. Places, facilities, artifacts, living entities (i.e., persons, plants, and animals), events, and even intangible cultural heritage items that may be placed on a map are deemed as POIs. Parts of the corpus presented here are also used to create a dedicated platform for use in secondary and higher education. Finally, the textual part of the Mythotopia corpus, comprising of parallel material in Greek and English, will be integrated into APOLLONIS, the Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation as a POS-tagged bilingual parallel corpus with open access. In particular, the corpus will enrich that part of the infrastructure related to language, namely, clarin-el, which serves as a virtual repository of language tools and resources forming web services in a stable, interoperable context (Gavrilidou et al., 2011).

3. Related work

Over the past decades, there has been a growing body of work aimed at the development, annotation, and processing of historical corpora in the humanities domain. In this respect, treebanks (that is, corpora annotated primarily at the level of syntax) have been developed for historical languages and language varieties, as for example Ancient Greek, Latin, Medieval English, and French – to name but a few. The treebank formalism allows representing sentences as tree-shaped graphs with nodes corresponding to the tokens (words) of the texts and edges (or arcs) corresponding to the syntactic information in terms of constituency (phrase structure) or dependency (grammatical relations). Besides syntax, treebanks may encode various other levels of linguistic analysis, from morphosyntax and lemma to semantics, and pragmatics. These datasets have been reported to pave the way for the analysis of historical languages by using state-of-the art Corpus Linguistics techniques for modelling various phenomena, for aiding literary and historiographical analysis, or for contributing to classification problems such as author attribution. Being a ground-breaking initiative of its kind, the Ancient Greek and Latin Dependency Treebank (AGLDT) comprises texts from the Perseus Digital Library coupled with syntactic, morphological, and lexical information (Bamman and Crane, 2011). At the same time, based on the corpus of the Index Thomisticus by Father Roberto Busa, the Index Thomisticus Treebank (IT-TB) is a dependency-based treebank which includes the opera omnia of Thomas Aquinas (Passarotti, 2019). Currently, the IT-TB is the largest Latin treebank available, whereas, besides morphosyntax, part of the data has also been annotated for pragmatics and semantics (tectogrammatical layer). The PROIEL family of treebanks (Eckhoff et al., 2018) comprises annotated corpora of early attestations of Indo-European languages, including Ancient Greek and Latin. The afore-mentioned initiatives seek to address scholarly as well as pedagogical concerns and/or make use of treebanks as a powerful resource for data-driven linguistic research, which is likely to have a great impact on the way the grammar of ancient languages is studied.

Beyond scholarly study, Language Technologies offer new ways of accessing cultural content in digital libraries (Bontcheva et al., 2002; Borin et al., 2007; Giouli et al., 2009); in this respect, benchmarks in the form of harmonised and standardised annotated corpora at least at the level of part-of-speech tagging and lemmatisation, as well as computational lexica have been proposed as minimum requirements for processing cultural heritage data (Borin et al., 2010; Giouli, 2011). Dependency annotation and annotation of Named Entities focused on toponym identification has also been successfully applied (Bontcheva et al., 2002; Nissim et al., 2004; Borin et al., 2007; Giouli et al., 2011) on literary and historical texts. One step further, novel types of annotations offer new ways of experiencing and visualising historical literary texts, via sentiment analysis (Jannidis et al., 2016), sensory (olfactory) information extraction (Tonelli and Menini, 2021), geotagging (Kew et al., 2019) and their combination (Heuser et al., 2016).

The relatively newly defined task of geotagging, also known as geolocation or location tagging, is aimed at assigning geographical coordinates (in the form of metadata) to input data (text, video, photographs, etc.). Geotagging approaches for estimating the locations of textual data (mainly Wikipedia articles or Twitter tweets) make use of Natural Language Processing techniques and language models (Brunsting et al., 2016; Kordopatis-Zilos et al., 2017) *inter alii*. Nevertheless, geotagging has been extensively studied in the context of multimedia resources and, in particular, with respect to geotagging photos in social network applications. Mu et al. (2020) provide a thorough review of the research literature, while Wong et al. (2017) focus on geotagging in tourism. Several social network-derived datasets can be exploited in geotagging research, including Flickr and Foursquare, although these datasets have not been created exclusively for this purpose. Sergieh et al. (2014) have created a quite large image dataset (~14 million geotagged photos) for automatic image annotation and reverse geotagging. The Yelp dataset (Yelp, Inc., 2011) provides a set of ~200 thousand POIs, accompanied by rich metadata, including coordinates, descriptions, categories, ratings, user reviews and photos, rendering it an ideal dataset for researching geotagging on multimodal data.

4. Corpus description

The main objective of the Mythotopia project is the systematic collection of texts from Greek and Latin literature that feature myths and mythological figures from the area of Eastern Macedonia and Thrace in Northern Greece, and the integration of the relevant literary and archaeological evidence in a database. Therefore, the mythological component lies in the heart of our endeavour, serving as the starting point that connects text, images, and video content that pertain to the domain of Archaeology and Travel. In this sense, the corpus comprises three sub-corpora depending on the domain or subject matter they belong to, namely: a sub-corpus of literary texts depicting the myths of the area, a multimodal corpus of digital combined representations of artifacts relevant to the myths, and a multimodal corpus in the domain of Travel and Tourism. Each domain-specific corpus may include text, images, or videos. In the following sections, we will elaborate further on the relevant components of the resource (sub-corpora).

4.1. The sub-corpus of literary texts

Myths are narratives that typically emerge in oral societies and seek to either teach (didactic) or explain the origins of things (aetiological). Greek mythology is centred around deities, heroes, and mythological creatures. Moreover, myths are linked to specific localities (places), religious or ritual practices and the culture of the time. In most cases, the stories of Greek mythology were initially shaped through poetic (or epic) tradition, before being disseminated in writing by means of the works of Greek and Latin literary authors. Therefore, the starting point of the mythological component of the project (and the relevant sub-corpus) consists of ancient Greek and Latin literary texts featuring myths or mythological figures and localities that are related to the area. These texts are excerpts selected from established scholarly editions; they cover a variety of genres and a wide range of Ancient Greek and Latin literary production both in prose (historiography, myth-writing, biography, rhetoric, philosophy, and scientific texts, such as geographical works and ancient scholia) and in verse (epic, drama, elegy, epigram, lyric poetry, bucolic poetry, and didactic poetry). Moreover, the sub-corpus at hand encompasses texts that pertain to the genre of travel literature, that is, travel writings produced by authors who visited the area of Eastern Macedonia and Thrace in antiquity (i.e., Pausanias, Strabo), and in Medieval and Modern times (i.e., Buondelmonti, Ciriaco d'Ancona). The Ancient Greek and Latin literary texts are stored in the original; consequently, their translations in Modern Greek (EL) and English (EN) are also provided for by project participants. Additionally, accompanying material (texts) in the form of biographies of authors, and general information about the mythological figures is also developed by project participants and included in the raw data. Finally, the sub-corpus comprises narrative texts in EL and their translations in EN that depict the myths that are relevant to the broader area of East Macedonia and Thrace. To date, 274 texts have been collected/produced amounting to c. 28,380 tokens.[1]

4.2. The cultural component

Besides literary texts, ancient Thracian myths have been a source of inspiration for both tangible and intangible cultural artifacts. As a matter of fact, pictorial representations of mythological scenes or imaginary figures of deities and mythological creatures are depicted on pottery, paintings, coins, and other artifacts of archaeological significance. Along the same lines, opera and Greek mythology have been intertwined for as long as opera itself has existed.

In this regard, the cultural component of the resource is a multimodal corpus that consists of images and texts relative to the myths and figures selected. More precisely, the collection encompasses images of artifacts dating back to Ancient Greek and Roman antiquity, as for example clay pots, sculptures, engravings, coins, various metal and glass objects, mosaics, sarcophagi, etc.

Finally, the sub-corpus comprises texts in EL which briefly describe the scene(s) and person(s) portrayed on the selected artifacts. In other words, they are verbal representations of the images. These texts have been created for the purposes of the current project and translated in EL. This is a rather heterogeneous component of the corpus and currently amounts to 48 artifact image files accompanied by explanatory texts of 3,583 tokens to date.

4.3. The sub-corpus in the domain of Travel

Finally, the corpus comprises textual data in EL and their translations in EN that pertain to the domain of Travel. These are primarily informative texts tailored to meet the needs of tourists visiting the area. In terms of content or subject matter, these texts feature primarily entities of the following types: living entities (i.e., animals, plants that are endemic in the area), geopolitical entities (i.e., cities, towns, villages, or minor settlements), geographical entities (e.g., mountains, rivers, beaches, lakes), facilities and archaeological sites, architectural and religious monuments, events and activities, as well as intangible cultural elements (i.e., food and gastronomy, folklore, and cultural events of the area). In essence, these entities constitute the body of POIs of the resulting application.

This sub-corpus comprises original material that was created either in the context of the current project or by experts in the specific domains; the latter were collected from various sources over the web. This results to further diversity of the content, which consists of both pre-existing and tailor-made material.

Apart from the primary textual data, images, and videos in the domain of Travel featuring the above-mentioned topics/subject matters were also collected from various sources over the web. To date, the sub-corpus in the domain of Travel consists of c. 1,200 texts (c. 13,066 tokens) and 80 images.

5. Methodology adopted for corpus creation

Developing a multimodal corpus that is so heterogeneous in terms of subject fields, modalities, and languages is far from being an easy task. Corpus creation was performed manually by experts on the various domains/subject fields. After the initial corpus creation specifications were set, the research teams in charge of creating and collecting the primary data were involved in researching and road mapping the available resources (myths, mythical figures, localities, authors, and corresponding literary texts available, POIs, etc.). Based on the above, a set of controlled vocabularies for data classification and indexing was defined; these vocabularies were consolidated early enough in the life cycle of the project and guided the development of the database schema thereof.

Corpus development involves the manual collection of literary texts and the selection of images based on the specifications set. Additionally, texts that pertain to the cultural and travel components of the corpus were developed from scratch; according to the guidelines set, these texts take the end-user into account. At the next stage, translations of the primary data, that is, the literary texts, and of the cultural and travel texts were provided. All the data (texts, images) were documented via the DB back-end. Currently, the linguistic processing of the textual data is underway. During corpus development all possible IPR issues were considered.

5.1. Building the corpus: selection criteria

Corpus development was performed incrementally using the myths of the area as the starting point and based on certain criteria. To this end, a set of specifications were defined to guide the development, formatting and documentation of the data that constitute the multimodal corpus.

As far as the mythological content is concerned, the main criterion was the level of relevance of the material to the region of Eastern Macedonia and Thrace (through a geographical place, a mythological person, etc.), followed by the existence of surviving texts with references. This second criterion was considered to ensure that proper documentation accompanies every myth. Finally, a more practical criterion was the amount of the existing archaeological material that also accompanies every myth that is included in the corpus.

Multimodal data were collected by experts in the field depending once again on the level of relevance to each of the selected myths. However, as the team found that there was an abundance of archaeological material available, it was decided that only the most representative artifacts were to be included in this component. Although a considerable part of existing material was not included, it was assured that the corpus contained a diverse variety of different items, be them statues, vases, reliefs, mosaics, coins, etc.

[1] Please, note that all numerical values relating to text sub-corpora represent the original content in Greek.

Finally, the sub-corpus pertaining to the domain of Travel mainly contains informative material and is, therefore, more inclusive in terms of scope; in other words, its components are not necessarily connected to the other sub-corpora or, for that matter, the myths that are included in the corpus. The major difference between this sub-corpus and the rest of the material is that this is not selected but rather produced originally for the purposes of this project. The main criterion of selecting the specific topics of the textual part of this sub-corpus was ensuring a wide coverage across geographical areas and the subject matters mentioned in section 4.3, whereas a decent level of uniform appearance guided the selection of the accompanying images.

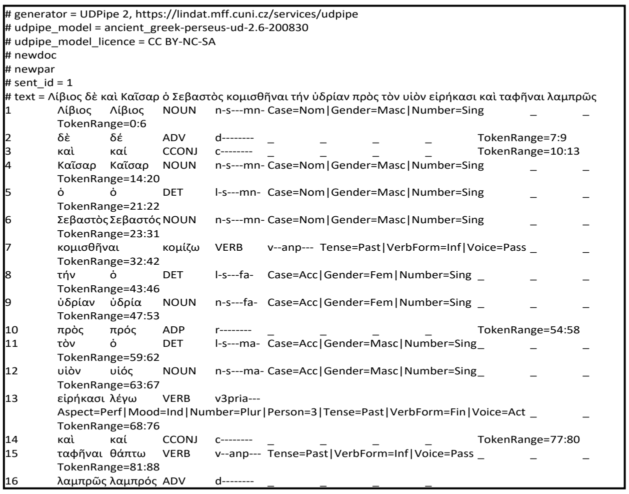
Furthermore, a booklet containing detailed principles for the selection, writing, editing, structure, form, documentation, and presentation of this sub-corpus was drawn up ensuring that all experts involved follow the same guidelines throughout the process.

5.2. Processing and adding metadata

Metadata have always been a key component of digital collections in the humanities domain aimed at organising and managing digital cultural and historical objects. In our case, metadata were added to both primary data and accompanying material (texts, images, video). Annotations were integrated across the following pillars: (a) efficient documentation aimed at indexing and retrieval of the content; (b) interlinking of the various entries in the database; (c) placing certain entities on the map, and (d) modelling linguistic features of the textual data. The latter was deemed necessary in view of not only providing a workbench for prospective developers but also integrating the final corpus into APOLLONIS infrastructure.

More precisely, documentation of the *literary textual* data involves the deployment of metadata elements which encode information necessary for text indexing, namely text title, author name, creation date (where applicable), language or language variety (Greek vs Ancient Greek, Latin, and English), etc. Additionally, information regarding text type/genre and topic was also added. Furthermore, to ensure documentation completeness and facilitate the inter-relation among primary textual data and the accompanying material (images, video, relevant texts), the documentation scheme has been extended accordingly to include these elements.

Documentation of the accompanying textual material deploys an upper-level hierarchically organized taxonomy that depicts the datatypes stored in the inventory. The taxonomy covers the domains of Archaeology, Literature, Mythology, and Travel, whereas the final model is compatible with existing standards, namely, the Text Encoding Initiative[2] and the upper-level ontology of selected sections of the CIDOC Conceptual Reference Model (CRM)[3]. More precisely, the metadata schema applied to texts that constitute the sub-corpus in the domain of culture includes the following elements: title/name of the artifact described in the text, data type, name of the artist/creator, creation date, place of origin, current location of the artifact, etc. Texts in the domain of Travel are also documented and indexed; additionally, they are geotagged. We will elaborate further on the process of geo-tagging in section 6. In most cases, the attributes to the aforementioned elements are entries of controlled and structured vocabularies that were created early in the life cycle of the project. Similarly, metadata were added to the images across the following axes: title, or caption of the image, the name of the creator of the image, the owner of the image (if different), and a url of the image itself. Moreover, annotation caters for interlinking among the various entities stored in the database. In essence, multiple associations between one or more myths and mythological persons documented in the project's database are retained so that visualisations of scenes and people or localities involved are feasible.



```
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe
# udpipe_model = ancient_greek-perseus-ud-2.6-200830
# udpipe_model_licence = CC BY-NC-SA
# newdoc
# newpar
# sent_id = 1
# text = Λίβιος δὲ καὶ Καῖσαρ ὁ Σεβαστὸς κομισθῆναι τήν ὑδρίαν πρὸς τὸν υἱὸν εἰρήκασι καὶ ταφῆναι λαμπρῶς
1       Λίβιος    Λίβιος    NOUN    n-s---mn- Case=Nom|Gender=Masc|Number=Sing          _       _
        TokenRange=0:6
2       δὲ        δέ        ADV     d--------  _       _       _       _       TokenRange=7:9
3       καὶ       καί       CCONJ   c--------  _       _       _       _       TokenRange=10:13
4       Καῖσαρ    Καῖσαρ    NOUN    n-s---mn- Case=Nom|Gender=Masc|Number=Sing          _       _
        TokenRange=14:20
5       ὁ         ὁ         DET     l-s---mn-  Case=Nom|Gender=Masc|Number=Sing         _       _
        TokenRange=21:22
6       Σεβαστὸς Σεβαστός NOUN    n-s---mn- Case=Nom|Gender=Masc|Number=Sing          _       _
        TokenRange=23:31
7       κομισθῆναι          κομίζω    VERB    v--anp---  Tense=Past|VerbForm=Inf|Voice=Pass _       _
        TokenRange=32:42
8       τήν       ὁ         DET     l-s---fa-  Case=Acc|Gender=Fem|Number=Sing _       _       _
        TokenRange=43:46
9       ὑδρίαν    ὑδρία     NOUN    n-s---fa-  Case=Acc|Gender=Fem|Number=Sing _       _       _
        TokenRange=47:53
10      πρὸς      πρός      ADP     r--------  _       _       _       _       TokenRange=54:58
11      τὸν       ὁ         DET     l-s---ma-  Case=Acc|Gender=Masc|Number=Sing_       _       _
        TokenRange=59:62
12      υἱὸν      υἱός      NOUN    n-s---ma- Case=Acc|Gender=Masc|Number=Sing_       _       _
        TokenRange=63:67
13      εἰρήκασι λέγω      VERB    v3pria---
        Aspect=Perf|Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin|Voice=Act _       _
        TokenRange=68:76
14      καὶ       καί       CCONJ   c--------  _       _       _       _       TokenRange=77:80
15      ταφῆναι   θάπτω     VERB    v--anp--- Tense=Past|VerbForm=Inf|Voice=Pass _       _       _
        TokenRange=81:88
16      λαμπρῶς λαμπρός ADV     d--------  _       _       _       _
```

Figure 1: A sample of the linguistic annotations in CoNLL-U format

Annotations relevant to pillars (a), (b), and (c) were added manually via a dedicated platform that serves as the front-end of the final database in which the final resource is being integrated (see section 5.3 below). The major advantage of manual annotation and indexing was the accuracy of metadata, which was ensured further by means of a collaborative or peer-review process involving humanities experts. On the contrary, the linguistic annotation of the textual data, which is currently underway, is being performed automatically. After extensive experimentation with our data, different tools were chosen per language. Part-of-speech tagging, and lemmatization of the Ancient Greek and Latin literary texts is feasible via the UDPipe pipeline (Straka et al., 2019), a language-agnostic tool trained on the latest Universal Dependencies 2.6 language models (Zeman et al., 2020); for our purposes, we chose the models trained on PROIEL 2.6 treebanks. The tool described in (Papageorgiou et al., 2000) was chosen for part-of-speech tagging and lemmatisation of Modern Greek texts. Currently, annotations are being performed automatically without further manual validation of the output since the tools employed have been reported to have an accuracy rate of about 96% (*ibid*, 2000). Following standard recommendations, all annotations are saved in CoNLL-U format. Annotations are encoded in plain text files (UTF-8), with three types of lines: word lines containing the annotation of a word/token, blank lines marking sentence boundaries, and comment lines starting with hash (#). Sentences consist of one (or more) word lines, and word lines contain the following tab-separated

[2] https://tei-c.org/

[3] https://www.cidoc-crm.org/Version/version-7.2

fields: the number of the word (ID) within the sentence, the word form, the lemma of the word form, the Universal part-of-speech tag, the language-specific part-of-speech tag, and the list of morphological features. A sample annotated sentence is shown in (Fig. 1).

5.3. Storing in a database

Documentation, interlinking, and geotagging of the corpus were performed via a platform that was ultimately used for storing and searching the corpus. On selecting the best foundation for our Content Management System (CMS), our primary concerns were wide usage, maturity, and continuity of the platforms used, while open source was another major focus. All the above combined also lead to security and minimal maintenance cost of our own platform. Thus, PHP was chosen on the backend, a Bootstrap derived CSS framework for styling, the Leaflet JS library for geocoding and other well-known utilities for user input validation, ease of access and interactivity. We opted for a purpose-built CMS over an established one. As a matter of fact, already existing CMS platforms also fully satisfy our concerns mentioned above. However, much of their complexity is not required in our use case.

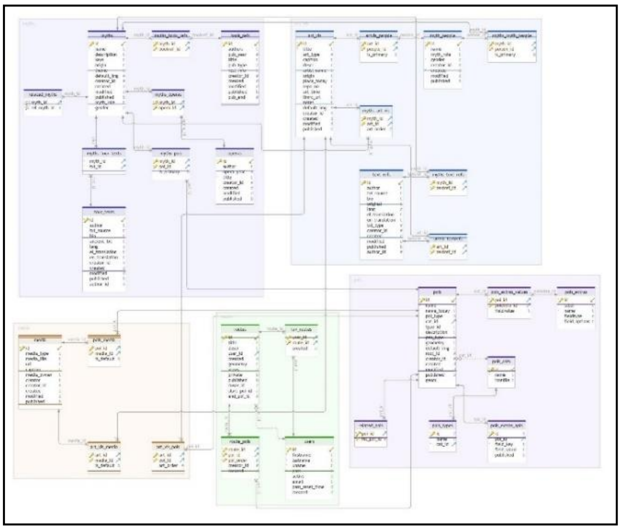
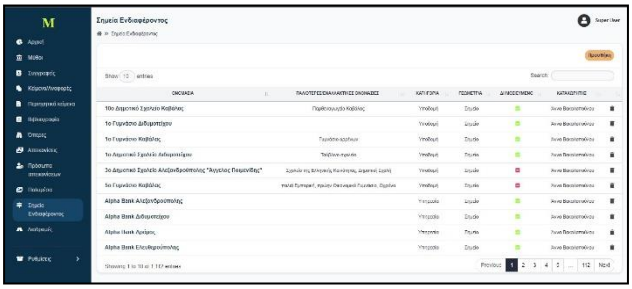


Figure 1: The main database schema

Moreover, some of the functionalities we require are provided through third party plugins, which do not necessarily share the same integrity. Thus, we felt safer developing our own system based on well-founded tools, rather than committing to, following, and evaluating several ready-made solutions. By doing so, we got significant gains in speed and resource utilization.



Σημεία Ενδιαφέροντος

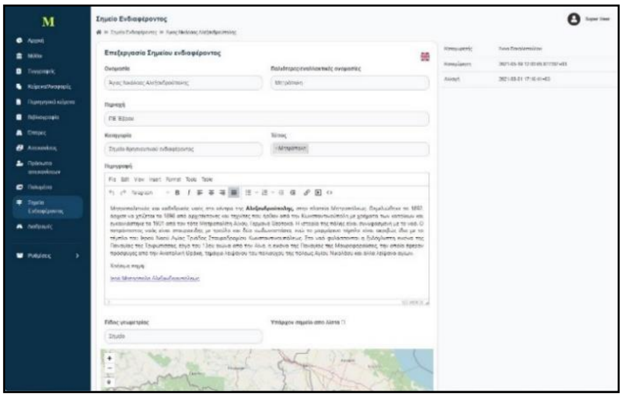
Show 10 entries

ΟΝΟΜΑΣΙΑ	ΠΑΛΑΙΟΤΕΡΕΣ/ΕΝΑΛΛΑΚΤΙΚΕΣ ΟΝΟΜΑΣΙΕΣ	ΚΑΤΗΓΟΡΙΑ	ΓΕΩΜΕΤΡΙΑ	ΔΗΜΟΣΙΕΥΜΕΝΟ	ΚΑΤΑΧΩΡΗΤΗΣ
10ο Δημοτικό Σχολείο Καβάλας	Παρθεναγωγείο Καβάλας	Υποδομή	Σημείο	🟩	[illegible]
1ο Γυμνάσιο Διδυμοτείχου		Υποδομή	Σημείο	🟩	[illegible]
1ο Γυμνάσιο Καβάλας	Γυμνάσιο αρρένων	Υποδομή	Σημείο	🟩	[illegible]
1ο Δημοτικό Σχολείο Διδυμοτείχου	Ταϊβάνειο σχολείο	Υποδομή	Σημείο	🟩	[illegible]
3ο Δημοτικό Σχολείο Αλεξανδρούπολης "Άγγελος Ποιμενίδης"	[illegible]	Υποδομή	Σημείο	🟥	[illegible]
5ο Γυμνάσιο Καβάλας	[illegible]	Υποδομή	Σημείο	🟥	[illegible]
Alpha Bank Αλεξανδρούπολης		Υπηρεσία	Σημείο	🟩	[illegible]
Alpha Bank Διδυμοτείχου		Υπηρεσία	Σημείο	🟩	[illegible]
Alpha Bank Δράμας		Υπηρεσία	Σημείο	🟩	[illegible]
Alpha Bank Ελευθερούπολης		Υπηρεσία	Σημείο	🟩	[illegible]

Showing 1 to 10 of 1,117 entries

Figure 2: Table representation of content for the Points of Interest

The platform follows the database schema and provides a menu containing the main data elements of the corpus: myths, reference texts, artistic representations, mythological figures (people), multimedia files and POIs. For each of these entities, the platform provides a common management mechanism. A table representation of already stored elements (Fig. 3), a form for editing purposes (Fig. 4), and a similar but empty form for data entry purposes. The use of this straight model of functionality provides easy understanding of the platforms' use, since project members who were responsible for storing and documenting the multimodal content, lacked adequate experience in content management applications.

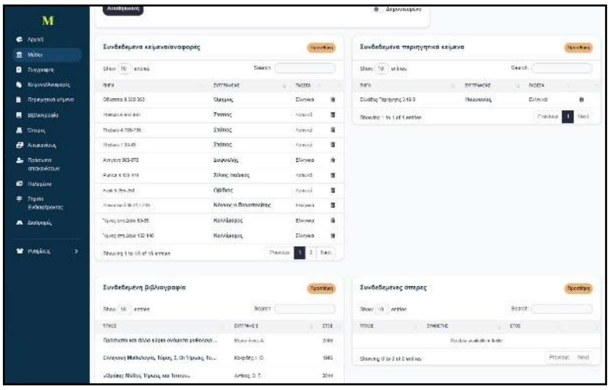


Σημείο Ενδιαφέροντος

Επεξεργασία Σημείου ενδιαφέροντος

Figure 3: Editing form for a stored point of interest.

All forms consist of several input text fields, some WYSIWYG editor fields, in cases we need formatted text, for all metadata applied to the selected corpus element. Interconnections with another element were provided using enriched dropdown elements. In the case of POIs, we also included a geo-tagging functionality which will be described in the next section. In some cases, where interconnections required a one-to-many relation and the nature of related elements was not easily expressed by a single title, we used separate table areas, providing more than one field of related metadata (Fig. 5).



Συνδεδεμένα κείμενα/αναφοράς

Συνδεδεμένα περιηγητικά κείμενα

Συνδεδεμένη βιβλιογραφία

Συνδεδεμένες όπερες

Showing 1 to 10 of 16 entries

Showing 1 to 1 of 1 entries

Figure 5: Interconnection of one-to-many corpus elements for a myth.

For security purposes and establishing a content quality control, our CMS uses a two-level user hierarchical system. At the lower level, annotators are granted access to all corpus elements and can create, edit, and delete their own records. Editors, on the other hand, have the same access to system functionalities, but also the right to publish completed records. After publishing, the record is no longer available for further editing access. On the top level, the Administrator has full access to all material, regardless of ownership or publishing status, and access to all user account managerial functionalities, needed in a CMS.

To facilitate complex relationships between entities, intermediate tables were used. Those consist merely of foreign keys to the referred entities. No distinct primary key is used on them, instead their primary key is composed of foreign keys only. This practically transforms them into an index. It ensures that no duplicates are inserted as this is a violation of the primary key. Furthermore, by utilizing constrains on the foreign keys we maintain the integrity of the database in case one of the connected entities is deleted. I.e., most modern database systems support the "on delete cascade" constraint on foreign keys. This, in combination with the usage of intermediate tables allows us to automatically remove all relationships, upon deletion, without affecting the referred entities and without the need to manually trace all relations.

6. Geo-tagging: placing the data on the map

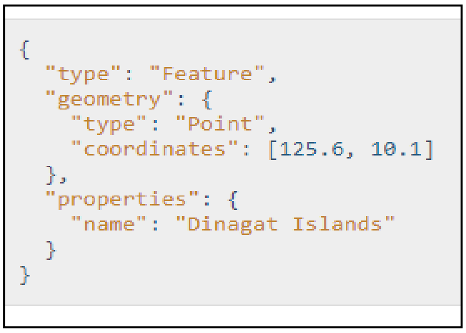
Besides the documentation of the cultural content, and the linguistic annotation added, one novel feature of the corpus is the fact that part of the textual data has been geotagged manually. More precisely, textual data that pertain to the domain of Travel, that is, POI entries, are assigned geographical coordinates.

In our case, to represent the geographic data stored in the Mythotopia corpus, OpenStreetMap was used. OpenStreetMap allows distribution of free geospatial data that can be used by anyone. We also used Leaflet (https://leafletjs.com/), which is a light, open-source JavaScript library that works efficiently across all major desktop and mobile platforms.

Annotation at this level is quite straightforward and no other expertise is required besides familiarity with the CMS. From there, annotators can easily depict a specific place on the map in two ways:

- By interacting with the map and clicking on the corresponding place. Annotators can adjust the zoom of the map as well as pan to the area on which they intend to place the marker. They can also delete the marker and place a new one in case of mistake.
- By using the "Forward & Reverse Geocoding" API[4], based on OpenStreetMap data. The specific API converts addresses into geocoordinates and vice versa: it uses reverse geocoding to cast geographic coordinates into human-readable addresses or toponyms. In the latter case, the annotators can type the coordinates into the corresponding text field and the API returns the name of the place - if found. The annotator can then click on the option provided and the details of the place are shown. If this is the correct place, the annotator confirms the choice, and a marker is shown on the map. Alternatively, annotators can start typing the name of the place they wish to depict on the map, and the API returns a list of places found. The annotator can then choose one of the places shown and confirm its coordinates. Then the corresponding marker is shown on the map.

In both cases, after the marker shown on the map is saved, the geographic data are successfully stored in the database in GeoJson format. GeoJson is a format that is used for encoding a variety of geographic data structures using JavaScript Object Notation (JSON). Its features may include point, line strings, polygon, circle, or multipart collections of previously mentioned features.



```
{
  "type": "Feature",
  "geometry": {
    "type": "Point",
    "coordinates": [125.6, 10.1]
  },
  "properties": {
    "name": "Dinagat Islands"
  }
}
```

Figure 6: GeoJson format example

Beyond the aforementioned functionalities for populating and geo-tagging the Mythotopia corpus, our work includes developing methods for semi-automatic enrichment of existing geospatial resources (POIs) with third party, open datasets. In particular, we are currently leveraging our previous work on toponym interlinking (Giannopoulos et al., 2021), by adapting the initially proposed similarity functions and machine learning training features in order to perform effectively on Greek toponyms and POIs. Our goal is to exploit these methods to enrich the core, manually gathered POIs of the Mythotopia corpus with metadata for the same entities, identified in third data sources, as for example *Geonames* and *ToposText*. More precisely, *Geonames*[5] is a large, global toponym dataset that contains circa 36 thousand place names in Greece. For each toponym, Geonames provides an extensive list of alternate location names. Thus, it comprises a valuable resource to be interlinked with/integrated into our corpus, especially considering that several Greek places are represented by more than one names. Similarly, *ToposText*[6] is a dedicated collection of places, ancient texts, and named entities related to the Greek history and mythology. Interlinking our corpus with ToposText will allow us to enrich the former with new metadata on myths regarding, e.g., related entities such as persons, locations, events.

7. Creating routes

In sightseeing scenarios, specifying a travel itinerary is essential for satisfying the travelers' preferences (e.g., means of transport) and restrictions (e.g., time limitations). However, even in cases of visiting only a few sites, the task of travel planning is challenging, since in most cases the travellers are not familiar with the road network and the local public transportation (Bast et al., 2016; Taylor, Lim, and Chan, 2018; Lopes, Coelho, and Santos, 2016; Tenemaza et al., 2020, Kaffes et al., 2018).

[4] https://rapidapi.com/GeocodeSupport/api/forward-reverse-geocoding

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

[6] https://topostext.org/

Our system provides routing functionality enabling users to generate different types of travel itineraries based on several criteria. In a nutshell, the user provides a starting and an ending location, as well as a set of POIs which they wish to visit. The system generates a route which comprises road segments/paths that need to be followed by the user, to visit all the desired POIs in a nearly optimal order. To do so, the system integrates several transportation networks (e.g., road networks, railways, boat routes) to generate routes based on different means of transport. These networks are extracted from OpenStreetMap, which is an open-source dataset, widely used in location-based services. The implemented routing system generates routes based on the following criteria:

(a) Route optimization objective. It defines the criterion which is attempted to be optimized during the route computation. It allows two options: *Route Length* and *Traveling Time*. Route Length corresponds to the overall length of the route, measured in meters. So, in this mode, the route is the shortest path that passes over all desired POIs. Traveling Time corresponds to the time that the user is expected to spend to travel along the whole route. The time is calculated by considering the length of the route and the excepted traveling speeds in the different road segments of the underlying road network.

(b) Travel type. It defines the means of transport to be used in the route, allowing three options: *Driving*, *Walking*, and *Transit* (e.g., train, boat). Travel Type determines the transport network on top of which the route is computed. For example, the transport network used in the Walking type contains the streets and the paths that pedestrians can use, while highways are omitted.

(c) Visiting order mode. It allows the user to select whether they prefer: (i) to visit the desired POIs in a specific order, or (ii) to allow the algorithm to find a nearly optimal order of visitation, with respect to optimizing the *Route optimization objective (a)*.

(d) Approximate route mode. It allows to adjust whether the route computation algorithm focuses on performance (speed of computation) or optimal quality (finding the shortest/fastest route). Selecting the approximate route mode, the generated route may not be the optimal, with respect to the route optimization objective (although it will be close enough), but its computation will be performed very fast. The approximate route mode enables the generation of routes that involve large number of POIs in a very short time. This mode is a great solution for interactive applications where low response time is essential.

The routing functionality involves the combination of shortest path algorithms, e.g., Dijkstra (Dijkstra, 1959), multi-criteria optimizations, approximation methods, and techniques for efficient route generation. Our implementation is mainly based on the methods proposed in our previous work in (Kaffes et al., 2018), which allows efficient route planning over multiple POIs.

The main idea of the implemented method is the following. Initially, we identify an adequate subset of the transportation network that is guaranteed to contain, among others, the optimal route for the user's input. This way, parts of the road network are pruned, reducing the search space of the algorithm, and enabling the efficient route generation.

To identify this network's subset, optimal paths (i.e., paths that optimize the *Route optimization objective*) between all possible POI pairs are computed using the Dijkstra algorithm over the network. Then, the resulting paths are merged, constructing a new network, i.e., an overlay network. This procedure guarantees that the optimal path is included in the generated network. Using the new network, a modified Dijkstra algorithm that finds an optimal path for the desired POIs visitation is used, to return the final route to the end user.

8. Conclusions and future research

We have presented a dataset that comprises a multilingual and multimedia corpus in the humanities domain. The challenge of long-term preservation and maintenance of digital project outputs calls for adding adequate and efficient metadata. In the paper, we have described the multi-layered and multi-faceted annotation performed so far, focusing on the process of geotagging textual data. Annotations have been performed via a dedicated platform that we developed in-house. Moreover, to render the resource useful after the end of the project life cycle, the linguistic annotation of the textual data is being underway as required in view of making the corpus available via a language resource infrastructure.

Future work has already been planned towards populating the corpus with new data, mainly consisting of (a) more original content in the domain of travel, (b) English translations across sub-corpora, and (c) new multimodal material, consisting of images and videos. Moreover, we plan to develop a RESTful API (Fielding, 2000) that could provide information of the myths themselves as well as the metadata they include, directly from the corpus. Such a service could be consumed by many applications since it supports a variety of devices and operating systems (cross-platform). As planned, the entire bilingual corpus will be POS-tagged, whereas the manual validation of the automatic processing has already been planned; Ultimately, the corpus will be incorporated as a standalone language resource into the national APOLLONIS infrastructure for Digital Arts, Humanities and Language Research and Innovation. Frome there, the corpus will be freely downloadable in CoNLL-U format and may be used under an open licence (CC BY 4.0).

9. Acknowledgements

We acknowledge support of this work by the project “Mythological Routes in Eastern Macedonia and Thrace” (MIS 5047101) which is implemented under the Action “Regional Excellence”, funded by the Operational Programme “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

9. Bibliographical References

Amitay, E., Har’El, N., Sivan, R., and Soffer, A. (2004). Web-a-Where: Geotagging Web Content. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK, pages 273–280.

Bamman, D., and Crane, G. (2011). The Ancient Greek and Latin Dependency Treebank. In C. Sporleder, A. van Den Bosch; K. Zervanou (eds.) *Language Technology* *for Cultural Heritage*. Berlin and Heidelberg: Springer, pp. 79–98.

Bast H., Delling D., Goldberg A.V., Müller-Hannemann M., Pajor T., Sanders P., Wagner D., and Werneck R.F. (2016) Route Planning in Transportation Networks. *Algorithm Engineering.*

Bekiari, C., Bruseker, G., Doerr, M., Ore, C.-E., Stead, S., and Velios, A. (2021). *Definition of the CIDOC Conceptual Reference Model*. V7.2.

Bontcheva, K., Maynard, D., Cunningham, H., and Saggion, H. (2002). Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content. In *Proceedings of the sixth European Conference on Research and Advanced Technology for Digital Libraries*, pp. 613–625. Lecture Notes in Computer Science.

Borin, L., D. Kokkinakis, and L. J. Olsson. (2007). Naming the past: Named entity and animacy recognition in the 19th century Swedish literature. In Proceedings of the ACL Workshop: Language Technology for Cultural Heritage Data (LaTeCH.), 1–8. Prague: ACL.

Borin, L., M. Forsberg, and D. Kokkinakis. (2010). Diabase: Towards a diachronic BLARK in support of historical studies. In *Proceedings of* LREC 2010.

Brunsting, S., De Sterck, H., Dolman, R., and van Sprundel, T. (2016). GeoTextTagger: High-precision location tagging of textual documents using a natural language processing approach. arXiv preprint arXiv:1601.05893.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik. 1: 269–271. doi:10.1007/BF01386390. S2CID 123284777.*

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O.E., and Jøhndal, M. (2018). The PROIEL treebank family: a standard for early attestations of Indo-European languages. Language Resources and Evaluation (52), pp. 29–65. https://doi.org/10.1007/s10579-017-9388-5.

Fielding R., (2000). Architectural styles and the design of network-based software architectures, Ph.D. thesis, University of California.

Gavrilidou, M., Labropoulou, P. & Piperidis, S. (2011). National Report on Language Technology in Greece.

Giannopoulos G., Kaffes V., and Kostoulas G. (2020). Learning Advanced Similarities and Training Features for Toponym Interlinking. *In Proceedings of the 42nd European Conference on IR Research, ECIR 2020: 111-125.*

Giouli, V. (2011). Annotating corpora from various sources in the humanities domain. Shortcomings and issues. Journal for Language Technology and Computational Linguistics, (JLCL), 26 (2). Special issue on Annotating Corpora for Research in the Humanities, pp. 53-63

Giouli, V., Glaros, N., Simov, K., and Osenova, P. (2009). A web-enabled and speech-enhanced parallel corpus of Greek - Bulgarian cultural texts. In Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education. LaTeCH - SHELT&R 2009, Athens, Greece.

Giouli, V., Simov. K., and Osenova, P. (2011). A parallel Greek-Bulgarian corpus: a digital resource of the shared Cultural Heritage. In In C. Sporleder, A. van Den Bosch, K. Zervanou (eds.) *Language Technology for Cultural Heritage*. Berlin and Heidelberg: Springer, pp. 99-114. Springer.

Heuser, R., Algee-Hewitt, M.A., and Lockhart, A. (2016). Mapping the Emotions of London in Fiction, 1700–1900: A Crowdsourcing Experiment. In P. Murrieta-Flores, D. Cooper, and C. Donaldson (eds), *Literary Mapping in the Digital Age*. Farnham: Ashgate.

Jannidis, F., Reger, I., Zehe, A., Becker, M., Hettinger, L. and Hotho, A. (2016). Analyzing Features for the Detection of Happy Endings in German Novels. arXiv preprint arXiv:1611.09028.

Kaffes V., Belesiotis A., Skoutas D., Skiadopoulos S. (2018) Finding shortest keyword covering routes in road networks. In *Proceedings of the International Conference on Scientific and Statistical Database Management.*

Kew, T., Shaitarova, A., Meraner, I., Goldzycher, J., Clematide, S., and Volk, M. (2019). Geotagging a Diachronic Corpus of Alpine Texts: Comparing Distinct Approaches to Toponym Recognition. In *Proceedings of the Workshop on Language Technology for Digital Historical Archives* in conjunction with RANLP-2019, pp. 11–18, Varna, Bulgaria.

Kordopatis-Zilos, G., Papadopoulos, S., and Kompatsiaris, I. (2017). Geotagging Text Content with Language Models and Feature Mining. In *Proceedings of the IEEE*, 105(10), pp. 1971-1986.

Lopes R.B, Coelho T., Santos B.S., (2016) Visually Supporting Location and Routing Decisions in Tourist Trip Planning: An Exploratory Approach. In *Proceedings of the International Conference Information Visualisation.*

Mambrini, F. (2016). The Ancient Greek Dependency Treebank: Linguistic Annotation in a Teaching Environment. In: G. Bodard, and M. Romanello (eds.) *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement*, pp. 83–99. London: Ubiquity Press.

Mu, S., Piwowarczyk, M., Kutrzyński, M., Trawinski, B., and Telec, Z. (2020). Comparative Analysis of Selected Geotagging Methods. *In Intelligent Information and Database Systems, 12th Asian Conference, ACIIDS. 10.1007/978-981-15-3380-8_29.*

Nissim, M., Matheson, C., and Reid, J. Recognising Geographical Entities in Scottish Historical Documents. In *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR ACM 2004,* Sheffield, UK.

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, S. (2000). A unified tagging Architecture and its Application to Greek. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece.* European Language Resource Association (ELRA).

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. In M. Berti (Ed.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, pp. 299-320. Berlin, Boston: De Gruyter Saur.

Vacalopoulou, A., Mastrogianni, A., Michalopoulos, C., Tsiafaki, D., Michailidou, N., Mourthos, I. Botini, P., and Stainhaouer G. (2021). Mythological Itineraries Along the Western Silk Road: Finding Myths in Visits to Eastern Macedonia and Thrace Today. In *Silk Road Sustainable Tourism Development and Cultural*

Heritage. The University of Thessaloniki and European Interdisciplinary Silk Road Tourism Centre.

Sergieh, H.M., Watzinger, D., Huber, B., Döller, M., Egyed-Zsigmond, E., and Kosch, H. (2014). World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging. *MMSys '14.*

Straka, M., Straková, J., Hajič, J. (2019): Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. In: ArXiv.org Computing Research Repository, ISSN 2331-8422, 1904.02099.

Taylor K., Lim K.H., and Chan J. (2018). Travel Itinerary Recommendations with Must-see Points-of-Interest. *In Proceedings of the Web Conference (WWW).*

Tenemaza M., Luján-Mora S., Antonio A., Ramírez J., (2020) Improving Itinerary Recommendations for Tourists Through Metaheuristic Algorithms: An Optimization Proposal. *IEEE Access 8.*

Tonelli, S., and Menini, S. (2021). FrameNet-like Annotation of Olfactory Information in Texts. In *Proceedings of the fifth Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.*

Wong, E., Law, R. and Li, G. (2017). Reviewing Geotagging Research in Tourism. In R. Schegg and B. Stangl (eds.). *Information and Communication Technologies in Tourism 2017: Proceedings of the International Conference in Rome, Italy, pp. 43-58. Cham: Springer Verlag, 10.1007/978-3-319-51168-9_4.*

Yelp, Inc. (2011). Yelp Dataset. Retrieved from: https://www.yelp.com/dataset.

Zeman, D., et al., 2020, Universal Dependencies 2.6, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3226.

Internet Resources

http://www.geonames.org/
https://getbootstrap.com/
https://leafletjs.com/
https://rapidapi.com/GeocodeSupport/api/forward-reverse-geocoding
https://tei-c.org/
https://topostext.org/
https://www.cidoc-crm.org/Version/version-7.2
https://www.php.net/