The Clavius on the Web Project: Digitization, Annotation and Visualization of Early Modern Manuscripts

Irene Pedretti
Historical Archives of the Pontifical Gregorian University
Piazza della Pilotta, 4
Roma, Italy
i.pedretti@unigre.it

Angelo Del Grosso, Emiliano Giovannetti, Lorenzo Mancini, Silvia Piccini
Institute of Computational Linguistics "A. Zampolli", CNR
via Moruzzi 1, Pisa, Italy
[name].[surname]@ilc.cnr.it

Matteo Abrate, Angelica Lo Duca, Andrea Marchetti
Institute of Informatics and Telematics, CNR
via Moruzzi 1, Pisa, Italy
[name].[surname]@iit.cnr.it

ABSTRACT

This paper describes the full procedure adopted in the context of the Clavius on the Web project, which aims to help Web users to appraise the importance of specific manuscripts by going beyond their digital reproduction. The proposed approach is based on the multilayered explication of linguistic, lexical and semantic data representing the innermost nature of the analyzed manuscripts. The final purpose of the project is to gather and display the results of the three layers of analysis through interactive visualization techniques and export them as Linked Data. All the analyses rely on the XML/TEI encoding of the text, followed by a CTS-based tokenization. As a working example for this paper, the analysis of a portion of a manuscript provided by Historical Archives of the Pontifical Gregorian University will be illustrated. The text is a letter written in Latin and sent by Botvitus Nericius to Christophorus Clavius in 1598 from Madrid.

Categories and Subject Descriptors

J.5 [**Computer Applications**]: Arts and Humanities - *linguistics*, *literature*; I.2.4 [**Computing Methodologies**]: Knowledge Representation Formalisms and Methods - *Representation languages*, *Semantic networks*; I.2.7 [**Computing Methodologies**]: Natural Language Processing - *Language parsing and understanding*.

Keywords

Promotion of Cultural Heritage, Early Modern Manuscripts, Clavius, NLP for Latin, Lexica and Ontologies, Data Visualization, Linked Open Data.

1. INTRODUCTION

Over the last few years the amount of manuscripts available on the Web has been increasing, due to the development of new online platforms able to host them. However, most of them are limited to the consultation of the manuscripts that can just be read and downloaded, for example without providing any mechanism to establish relationships between cited entities or to investigate the domain specific terminology.

This paper describes a multilayered analytical procedure which makes it possible to access manuscripts and their content. The results of the various steps are integrated into a Web platform accessible for academic, scholastic or pleasure purposes. The user is able to explore the content of manuscript in depth, by finding out new information about cited people, places and events, to browse a lexicon and a domain ontology in order to learn new words and concepts and to better understand the language and the structure of sentences. Furthermore, the Web platform allows a user to examine the physical appearance of a manuscript in detail thanks to high resolution digitization processes.

In particular, the Web platform presents the manuscript along with three different layers of analysis: linguistic, lexical and semantic. The aim of the linguistic analysis is to associate a morphological reading and a lemma to each orthographic unit (token), which the text is composed of. The results of this analysis can be used i) for linguistic based searches (e.g. by lemma), ii) as a source of information to help scholars and students in the study of the language, and iii) as a basis for further NLP related tasks.

As far as the lexical level of analysis is concerned, a thesaurus-lexicon of Clavius' scientific terminology was built. Its primary purpose is to provide both the expert and the everyday user with help in gaining a deeper understanding of the issues addressed in the corpus of letters as well as the main knowledge in mathematics and astronomy during the Renaissance.

Finally, the semantic analysis provides a deeper insight into the content of the manuscripts by annotating relevant entities, such as historical characters, places, events, and bibliographical references. Relationships between entities are defined as well and each entity, when possible, is linked to a Named Entity Repository, which is exposed as a Linked Data endpoint, connected with other sources available on the Web such as DBpedia, VIAF and GeoNames. Integration between all the layers of analysis is achieved by means of Canonical Text Service URNs,[1] which allow a single token and a single annotation to be univocally identified.

The Web platform is part of the Clavius on the Web project,[2] which aims to foster and enhance the study of the manuscripts of

AIUCD'14, September 18-19, 2014, Bologna, Italy
 ISBN 978-1-4503-3295-8/14/09…$15.00
DOI: http://dx.doi.org/10.1145/2802612.2802636

[1] http://www.homermultitext.org/hmt-doc/cite/cts-urn-overview.html.

[2] http://www.claviusontheweb.it/.

the Historical Archives of the Pontifical Gregorian University (APUG). APUG owns more than 5,000 manuscripts related to the activity of the Jesuits of the Roman College (1551-1773). The Clavius on the Web project pays particular attention to those related to Christophorus Clavius (1538-1612), a Jesuit mathematician and astronomer, and one of the most respected and influential scholars of his time [1].

2. STATE OF THE ART

The Clavius on the Web Project is the result of a detailed study of the state-of-the-art solutions for general-purpose digital libraries and monothematic archives. Most initiatives focus on Web access to repositories of texts and images available through collaborative and interconnected frameworks. On the one hand, there are general purpose initiatives, the best example of which being: Gallica,[3] the Library of Congress,[4] as well as international projects such as Internet Archive,[5] Europeana[6] and Google Cultural Institute.[7] On the other hand, some of the main projects about single authors' works concern Van Gogh,[8] Wittgenstein[9] and Nietzsche.[10]

Platforms such as Knowledge Circulation in the 17th Century[11] and Darwin Correspondence Project[12] explore and analyze corpora of letters in innovative ways, similarly to our work. A significant transcription framework for unstudied manuscripts has been developed by the Transcribe Bentham project.[13] It is worth mentioning DARIAH,[14] TEXTGRID,[15] CLARIN[16] and Interedition[17] as research infrastructures for digital textual scholarship.

Two projects which influenced our work greatly are: The Perseus Digital Library,[18] the largest archive of Greek and Latin texts, and The Homer Multitext project,[19] a framework for digital philology concerning texts and facsimiles of the Iliad and the Odyssey.

More specifically, as will be described in the following sections, we have exploited the Latin treebank and the lexical database of the Perseus Project and the notation scheme of CTS architecture, developed in the context of The Homer Multitext project.

[3] http://gallica.bnf.fr/.

[4] http://www.loc.gov/library/llibarch-digital.html.

[5] http://archive.org/.

[6] http://www.europeana.eu/.

[7] http://www.google.com/culturalinstitute/.

[8] http://www.vangoghletters.org/.

[9] http://wittgensteinsource.org/.

[10] http://nietzschesource.org/.

[11] http://ckcc.huygens.knaw.nl/.

[12] https://www.darwinproject.ac.uk/.

[13] http://blogs.ucl.ac.uk/transcribe-bentham/.

[14] http://dariah.eu/.

[15] http://www.textgrid.de/.

[16] http://clarin.eu/.

[17] http://www.interediction.eu/.

[18] http://www.perseus.tufts.edu/.

[19] http://www.homermultitext.org/.

3. THE CHALLENGE OF CLAVIUS ON THE WEB

The objective of Clavius on the Web is, in a way, similar to other digital correspondence projects: to rebuild the network created by the letters written by people from different countries over a specific period of time. In addition to this main goal, the Project has faced the new challenge of getting into the texts to make the information of different informative layers (linguistic, lexical and semantic) explicit and machine actionable according to the Linked Data philosophy. In particular, we developed a Web platform which integrates annotation tools specifically aimed to extract textual elements and entities belonging to different informative layers: linguistic, lexical and semantic. Some of the annotated elements have been afterwards formally structured inside a lexicon, a domain ontology and a Named Entity Repository. The elements belonging to the different layers of analysis are presented to users through three distinct interactive visualizations designed for non-expert users, with the aim of reaching the widest possible dissemination of information.

However, the scholarly impact of this initiative is not to be considered of secondary importance, since it will allow Clavius' work to be studied in a new light as well as in more detail. In order to illustrate the full process of analysis, from the digitization of the manuscripts to the interactive visualization on the Web, a letter sent by Botwid of Närke to Christophorus Clavius will be presented as a case study. Since the letter is written in Latin, translations into Italian and English are also provided.

4. DIGITIZATION, TRANSCRIPTION AND TRANSLATION

Clavius' correspondence is mainly preserved in two manuscripts,[20] containing about 300 letters, sent to Clavius by 134 different people of 14 nationalities, located in 127 different places, during the years 1570-1612. Until 2012 these papers were in a bad state of conservation and, due to an old binding, it was very difficult to read them. In that year APUG restored the letters and started their digitization. A Nikon D3X camera set up with a focal length of 60 and a focal ratio of 2.8 was used to photograph them. These ones have a resolution of 300 dpi and are available in two formats, TIFF for storing and the compressed JPEG for use on the Web.

The transcriptions of the letters were initially done by Ugo Baldini and Pier Daniel Napolitani in their edition of Clavius' correspondence [2]. These transcriptions were revised, corrected and marked up using TEI P5[21] (see Section 5). An accurate physical description of the letters (including a study of watermarks, hands and other material aspects) was added in the TEI-Header.

Some of the letters written in Latin were translated into Italian and English to make them accessible to a wider international audience. The physical description, originally in Italian, was translated into English as well. The translations were marked in TEI, following the CTS structure at a sentence level, thanks to which it is possible to link a sentence to its translation (see Section 5).

[20] APUG 529 and APUG 530.

[21] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/.

To give an idea of the workflow of the project, a letter sent by Botwid of Närke (also known as Botvitus Nericius; about 1540-post 1599) to Clavius in 1598 will be used as an example.[22] Nericius, born in Sweden, began studying mathematics in Rome, probably under the teaching of Clavius [2]. As with the other four letters sent by Botwid to Clavius (1597-1599), this one was sent from Madrid, where Botwid was a member of the Academy of Mathematics, founded in 1582 by Philip the Second and initially inspired by the well-known architect and mathematician Juan de Herrera (1530-1597) [3].

The text, dated 12th September 1598, is the only one signed by Botwid among the letters sent to Clavius. Here the Swedish mathematician tries to give a more valid proof of the construction of the quadratics of Dinostratus (a plane curve used to solve the problem of the quadrature of the circle), particularly on the identification of its extreme point.

5. LETTER ENCODING: TEI XML

The text of Clavius' correspondence was encoded using the XML schema provided in the TEI version P5. The hierarchical structure of XML/TEI used to encode each letter is compliant with the CTS requirements. The CTS URN syntax is composed of an invariant[23] and a variant[24] part[25] that follow the hierarchical structure of the letter. Below the XML/TEI structure of a portion of the letter is shown:

```
<div type="letter" n="l_530-147">
  <div type="opener" n="o_01">
   <ab type="paragraph" n="ab_01" >
    <milestone unit="page" n="129r"/>
    <s type="sentence" n="s_01">
    <lb n="01"/>
       Botvitus Nericius Christophoro Clavio S.
    </s>
   </ab>
  </div>
  <div type="body_text" n="t_02">
   <ab type="paragraph" n="ab_02">
    <s type="sentence" n="s_02">
    <lb n="02"/>
       Occupationibus tuis Doctissime
       Clavi parcens,
    <lb n="03"/>
       in tertia mea ad te epistola,
       praesertim vero in
    [...]
    </s>
   </ab>
  </div>
</div>
```

The element with the attribute @type (e.g. letter, opener, body_text, closer) was chosen for the encoding of the structural parts of the letter instead of the elements <opener> and <closer>. The <milestone> tag, encoding the recto and verso of each page of the sheet, is placed at the top of each page to optimize the performance of the tokenization tool. The <ab> tag, encoding the text blocks, was chosen instead of <p> to keep the document structure as general as possible. The @type attribute defines the type of text block enclosed by <ab>, and it often corresponds to a paragraph. The sentence and its content are encoded with the <s> tag; the @type attribute makes it possible to specify the type of content (e.g. address, dateline). The line (<lb>) is the last encoded element of the hierarchical structure. Each element is also specialized by the @n attribute. The @n is used in order to identify unambiguously the elements and to allow the necessary references to the CTS notation schema. Usually, the value of @n is formed by an acronym identifying the TEI element to which it refers (e.g. l for letter; o for opener; t for body text; ab for text block; s for sentence; lb for line break), followed by an underscore (_) and a number (e.g. 530-147 demarks the letter identifier; strings as 01, 02, etc. represent progressive numbers).

6. TOKENIZATION AND CTS

The text of the letters, transcribed and encoded as described in Section 5, was automatically split into orthographical units, called tokens. A token, therefore, is a sequence of alphanumeric characters extracted from the text and it is interpreted as an independent and uniform processing unit [4], [5].

The following XML snippet describes the first and second token belonging to the second line of the sentence labelled as s_02:

```
<tokens>
  <token
    uri="urn:cts:histReSci:clavius.apug.edApug:
   l_530-147.t_02.ab_02.s_02@Occupationibus[1]"
   start="41"
    end="55">
      Occupationibus
  </token>
 <token
    uri="urn:cts:histReSci:clavius.apug.edApug:
   l_530-147.t_02.ab_02.s_02@tuis[1]"
    start="56"
    end="60">
      tuis
  </token>
</tokens>
```

Tokens[26] are delimited by white spaces or punctuation marks, which are tokens themselves. For example, the sequence of characters "secunda illius parte," contains four tokens: [secunda] [illius] [parte] [,]. Moreover, abbreviations are resolved, such as the token related to the following TEI chunk corresponding to [magister]:

```
<choice>
 <abbr>mag.r</abbr>
 <expan>magister</expan>
</choice>
```

Hyphenated words are normalized, as the token [Dinostrati] in the following XML snippet:

```
<w>Dino<lb n="05"/>strati</w>
```

[22] This letter has signature APUG 530, l. 129-130 and in Clavius, 1992 is the number 147.

[23] urn:cts:histRenSci:clavius.apug.edApug

[24] l-530-147.o-01.ab-01.s-01@Botwitus[1]

[25] The namespace @histRenSci@ identifies the domain of History of Science in the Renaissance; textgroup: clavius; work: apug; edition identifier: edApug; the translation is identified by the trIta for Italian and trEng for English.

[26] Up to now, any issues about sub-tokenization and/or multiword management have not yet been taken into account.

In conclusion, editorial expansions are adopted, e.g. the TEI addition

```
<w>sevi<ex>sti</ex></w>
```

is resolved in [sevisti].

The token identifier is represented in the CTS sub-reference notation [6]. For example, the token [epistola], extracted from the second sentence of the letter, has the following CTS URI:

```
urn:cts:histReSci:clavius.apug.edApug:
l_530-147.t_02.ab_02.s_02@epistola[1]
```

Finally, each token also contains offsets of its start and end positions in the text. The aforementioned tokenization phase is mandatory to decouple the textual source from the various levels of annotations, thus allowing each analysis process to refer to the same centralized resource. It is natural, therefore, to exploit a model providing "linked data" support and network/graph-oriented approaches.

7. LINGUISTIC ANALYSIS

The process of linguistic analysis is aimed at labelling each single token with linguistic information [4], [5]. A first prototype of a semi-automatic lemmatization tool for Latin was implemented.[27]

The tool was designed as a statistical PoS tagger supported by a lexicon and a forms-lemmata database. At the end of the classification process the output is manually checked through a proofreading system. In contrast, the well-known LemLat [7] and Morpheus [8] engines do not embed a disambiguation mechanism, thus providing every possible morphological reading for each word, sometimes producing erroneous results.

The linguistic analysis is carried out in a series of steps: first, the tool extracts the sentences from the encoded transcriptions on the basis of their CTS location-independent and shared identifier. Second, the tool associates morpho-syntactical information to each token of the sentence (see Section 6) by using the HunPos engine [9]. HunPos is one of the most used and efficient supervised PoS taggers: it is based on Hidden Markov Models and developed to manage languages with complex morphology. Finally, the lemmatization of each token (word) is carried out by querying a rule based lemmatizer [7] and the Perseus repository of word-forms and lemmata [10]. The Latin Dependency Tree-bank [11] was used to train the first prototype of the linguistic analyzer.

The training set was composed of 53.143 tokens, 3.474 sentences, and 438 distinct morphological readings. In addition, the prototype is supported by a proofreading component.

The results of the linguistic analysis are generated and stored in XML. The `<linguistic_analysis>` element is the root of the data structure. Furthermore, each sentence has as many children elements `<token>` as the tokens it is composed of. The attributes of the `<sentence>` element include (a) the `@uri` for recording CTS-URN identifier, (b) the `@start`, `@end`, and `@span` for storing the position and range information.

A token element bears attribute about (a) the CTS-URN identifier (`@uri`), (b) the progressive number (`@prog`), (c) the start and end character positions (`@start`, `@end`), (d) the string representing the word-form of the token (`@form`), (e) the morphological reading (`@morphoCode`), and (f) the word lemma (`@lemma`).

A snippet of the linguistic analysis XML is listed below:

```
<linguistic_analysis>
  <sentence
    uri="urn:cts:histReSci:clavius.apug.edApug:
          l_530-147.t_02.ab_02.s_02"
    start="41"
    end="326"
    span=
    "l_530-147.t_02.ab_02.s_02@Occupationibus[1] -
     l_530-147.t_02.ab_02.s_02@.[1]">
   <!-- [...] -->
   <token
     uri="urn:cts:histReSci:clavius.apug.edApug:
          l_530-47.t_02.ab_02.s_02@Doctissime[1]"
     prog="3"
     start="20"
     end="30"
     form="doctissime"
     morphoCode="a-s---mvs"
     lemma="doctus" />
  </sentence>
</linguistic_analysis>
```

8. LEXICAL ANALYSIS

Besides the linguistic analysis, an electronic thesaurus-lexicon of Clavius' mathematical-astronomical terminology was built alongside a *lexical ontology*. In particular, the lexicon is constituted by a set of *lexical entries*, each associated to one or more *lexical senses* via the `has_Lexical_Sense` relation. Each lexical sense is linked to a concept, formalized in a lexical ontology, via the `has_Ontological_Reference` relation. These terminological and ontological resources were both built and managed using the Protégé-OWL platform.[28]

The *lexicon* is populated by lexical entries (representing the lexemes) extracted both from Clavius' epistolary corpus and works on gnomonics, practical arithmetics, practical geometry and algebra, and from his commentary and translations of Sacrobosco's *De Sphaera mundi*, Euclid's *Elementa*, Theodosius of Bithynia's *Sphaericorum Libri*. For the first time the fundamental terms of the texts, on which scholars such as Réné Descartes and Marin Mersenne were formed [1], are modeled in this lexicon in a highly formal and structured representation inspired by SIMPLE [12], a well-known lexical model in Computational Lexicography. This model was first adopted for the creation of harmonized electronic lexica for twelve European languages; it is considered a *de facto* standard and has greatly inspired the Lexical Markup Framework, the ISO standard for Natural Language Processing lexica and Machine Readable Dictionaries. Thanks to its flexibility, it makes it possible to formalize lexical senses belonging to different domains of knowledge, at the granularity level deemed the most appropriate. According to the Generative Lexicon theory [13], on which basic assumptions the semantic representation is grounded, SIMPLE permits a highly structured representation even of those lexical entries characterized by a more complex semantic content.

[27] The Java source code is shared as CC BY- SA and located at the following repository: https://github.com/CoPhi/ClaviusLemmata.

[28] http://protege.stanford.edu/.

The lexical entries constituting the lexicon are described by morpho-syntactic distinctive traits (e.g. PoS). The lexical senses are formalized through paradigmatic relations (hyperonymy, hyponymy, meronymy, and holonymy) and with information on the domain of use and, whenever appropriate, on the type of event denoted. As far as the lexical ontology is concerned, orthogonal dimensions of meaning (telic or agentive), which are expressed through the Qualia Structure,[29] are used.

For the time being the Clavius on the Web lexical ontology consists of 31 classes, organized along 4 hierarchical levels.[30] A snippet of the OWL representation of the lexical sense of "`line`", relative to the term that occurs in the third sentence of the letter 147, is illustrated below:

```
<ClassAssertion>
    <Class IRI="#Geometric_Entity"/>
    <NamedIndividual IRI="#linea"/>
</ClassAssertion>
<ObjectPropertyAssertion>
    <ObjectProperty IRI="#isa"/>
    <NamedIndividual IRI="#linea"/>
    <NamedIndividual IRI="#magnitudo"/>
</ObjectPropertyAssertion>
```

Such a structured organization of lexical information, which emphasizes the componential and relational nature of word meaning, might suggest new paths of analysis and help to gain a wider knowledge of the overall domain terminology. All the lexical senses of the lexicon are linked to their occurrences in the letters trough a semi-automatic lexical annotation process: thanks to the previous linguistic annotation, all the inflected forms of a lemma can be identified in the text, and their correspondence to the lexeme encoded in the lexicon can be manually validated. The lexical annotation of the word "`line`", the formal representation of which was mentioned above, is shown below, using the CTS notation:

```
<lexical_analysis>
  <lexical_entity
    class="Geometric_Entity"
    NamedIndividual="#linea">
     <token
       uri="urn:cts:histReSci:clavius.apug.edApug:
       l_530-147.t_02.ab_02.s_03@lineae[2]" />
  </lexical_entity>
</lexical_analysis>
```

9. SEMANTIC ANALYSIS

Letters were annotated semantically, in order to recognize people, places etc. The semantic annotation was performed manually through an open source tool, called Brat.[31] To support the semantic annotation, a domain ontology was developed in parallel with the definition of the lexical ontology illustrated above. This ontology includes the entities represented by the following basic classes: `Person`, `Place`, `Letter`, `Time`, `Instrument`, `AstronomicalEntity`, `Work`, `Event` and `Group`.

Technically, it reuses and extends some standards, such as FOAF[32] for people, the GeoNames ontology[33] for places and bibo[34] for letters in order to link the entities to the Web of Data. For example, within the Project, the class Letter was enriched with the property `quotes` (and its inverse `is_quoted`) in order to describe entities quoted within a letter.

9.1 Named Entity Repository

All the entities related to the Clavius domain are contained in the Named Entity Repository (NE Repo). Letters, People and Places were manually inserted into the repository, while Dates, Groups and Instruments were extracted from the semantic annotation of letters. Each entity corresponding to a Person or a Place was enriched with information available on the Web (Wikipedia, VIAF, etc.) or in specific sources, such as Baldini [2]. The NE Repo provides both internal and external links. Internal links establish relationships between people, places and writings, while external links retrieve specific information about them from the Web.

A relationship between two entities can be one of the following: a) `quotes`: relates letters to cited entities, e.g., Letter 147 quotes Clavius (`is_quoted` being its inverse relation); b) `author_of/recipient_of`: denotes a given person as author or recipient of a given letter, e.g. Clavius is the recipient of letter 147; c) `place_of`: the place where a letter was written/received d) `born_in/died_in`: the place where a person was born or died.

The NE Repo constitutes a *knowledge graph* of the Clavius' domain. It is exposed as a SPARQL node[35] and is released under Creative Commons CC BY-SA.

9.2 Semantic Annotation

An example of semantic annotation is shown below using the CTS notation. Currently, the references to the NE repo are inserted as HTTP URLs through the attribute `@individual`:

```
<semantic_analysis>
  <entity object="urn:cts:histReSci:
   clavius.apug.edApug:l_530-147.
   o_01.ab_01.s_01@Botvitus[1]
     urn:cts:histReSci:clavius.apug.edApug:
      l_530-147.o_01.ab_01.s_01@Nericius[1]"
     class="Person" individual="Person:91"/>
  <entity object="urn:cts:histReSci:
     clavius.apug.edApug:l_530-147.
      o_01.ab_01.s_01@Christophoro[1]
     urn:cts:histReSci:clavius.apug.edApug:
      l_530-147.o_01.ab_01.s_01@Clavio[1]"
     class="Person"
     individual="Clavius:Person:135"/>
 </semantic_analysis>
```

where `Clavius:` stands for http://claviusontheweb.it/dataset/.

[29] According to Pustejovsky, the Qualia Structure is one of the four levels of semantic representation; it is composed of four roles: the constitutive role (denoting the relation between the entity and its constituents), the formal role (distinguishing the entity within a larger domain), the telic role (denoting the function of the entity) and the agentive role (indicating the origin of the entity).

[30] The ontology structure is constantly evolving: the encoding of new entries will allow for refining and extending of the present ontology in compliance with the building principles of the archetypal model.

[31] http://brat.nlplab.org/index.html.

[32] http://xmlns.com/foaf/spec/.

[33] http://www.geonames.org/ontology/documentation.html.

[34] http://bibliontology.com/.

[35] http://claviusontheweb.it/dataset/snorql/.

10. VISUALIZATION

Three HTML5 Web user interfaces are provided to browse all the aforementioned data: the first is focused on the manuscript, the second on the linguistic and semantic annotations, while the third displays an interactive view of Clavius on the Web's *knowledge graph*.

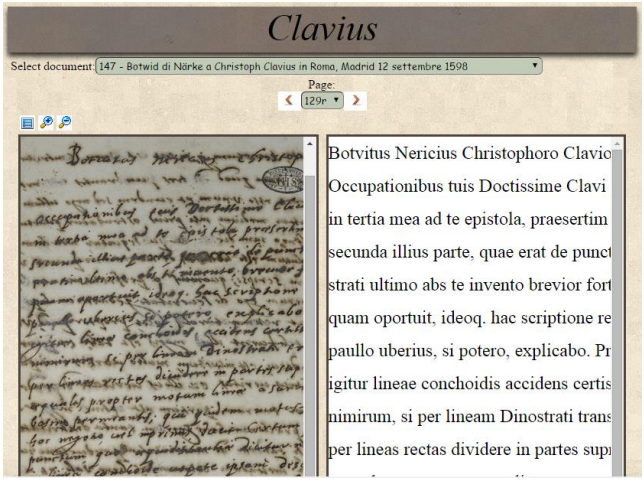

Figure 1. The manuscript interface, showing the first part of letter 147. Both the digital image and the transcription are displayed.

In the *manuscript interface* (Figure 1), users can see the digital image of the manuscript, and zoom in to appreciate its details. In the rightmost part of the interface, the user can also read its transcription. This fundamental visualization constitutes the starting point for users to access the other two, more specific, interfaces.

The *visualization of linguistic and semantic annotations* is conceived to be appealing to the public at a wide audience including, people with poor technical and/or domain expertise. Typography, symbols, layout and choice of color play a central role, like in an information graphic. All of these elements are combined in a readable display reminiscent of musical notation, e.g. each sentence appears as a musical measure.

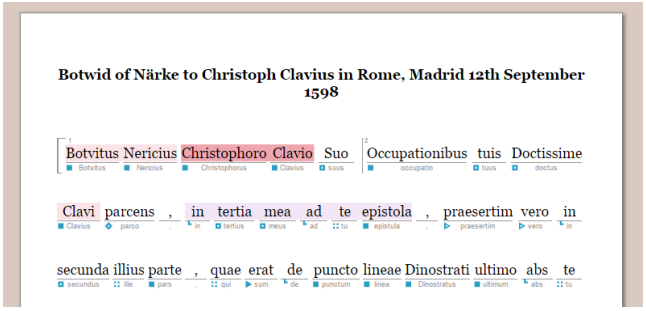

Figure 2. The visualization of linguistic and semantic annotations. Sentences, parts of speech, lemmata and semantic classes are displayed in a single depiction, while keeping the readability of the text.

Figure 2 illustrates the current visualization prototype, which highlights different aspects of the analysis. Sentence splitting is represented by means of a gray vertical line. Each token is underlined, with the associated lemma placed below it. The PoS is denoted by a blue symbol under the word, i.e. a full square for nouns, a full triangle for verbs, a hollow square for adjectives, a hollow triangle for adverbs, etc. Portions of text denoting entities are highlighted with different colors according to their class (e.g. red for People, purple for Works). Although conceived for presentational purposes, the visualization prototype was also successfully tested as a graphical proofreading tool. The interface appeared to be very useful for the revision process thanks to its capability of leveraging the user's pre-attentive processing [14], i.e. their ability to immediately perceive graphical features such as color, contrast or size.

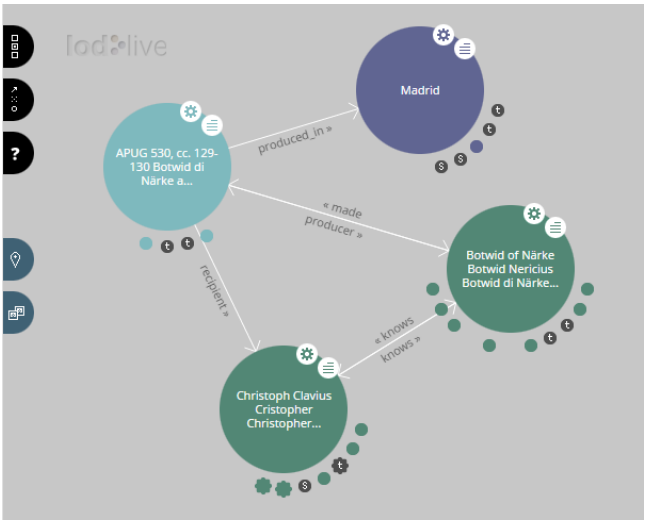

Figure 3. The knowledge graph visualization, using LodLive to show three of the entities connected to letter 147 (namely, the sender, the receiver and the place from which the letter was sent).

The third interface (Figure 3) uses LodLive [15], which allows for navigating the *knowledge graph* discussed in Section 9.

11. CONCLUSIONS AND FUTURE WORK

This paper introduced the approach we adopted for the Web publishing of the digital edition of Clavius' correspondence. The text of the manuscripts was enriched by a series of distinct layers of analysis integrated by means of Canonical Text Service URNs. The outcomes of the analysis are shown using *ad hoc* visualization techniques and exposed as Linked Data.

As future works we plan to improve the adopted technologies and add more features. Concerning the acquisition steps, we plan to develop a tool for the assisted transcription, translation and TEI encoding of the letters. Another improvement will focus on the linguistic analysis with the inclusion of domain adaptation techniques. As a matter of fact, the NLP models were trained using treebanks of classical Latin, while most of the project corpus is written in Renaissance scientific Latin. In addition, we will work on a proofreader to manually correct the automatic linguistic analysis. As a further step, we are also testing reasoning approaches to be applied to the NE Repo. Finally, about the enhancement of tools, we plan to add more layers to the visualization interfaces, e.g. morpho-syntactic traits, such as the Latin case, will be integrated in the design by using different colors and new forms of visual representation.

In parallel, additional manuscripts and letters by Clavius will be analyzed and integrated inside the knowledge graph, thus

providing more data for the activities of lexical and semantic annotation. The lexical and domain ontologies, as well as the NE Repo, will be consequently enriched with new lexical entries, classes, relations and entities.

ACKNOWLEDGMENTS

The Clavius on the Web Project was financially supported by Registro.it.

REFERENCES

[1] Lattis, J. 1994. *Between Copernicus and Galileo: Cristoph Clavius and the collapse of Ptolemaic cosmology*. The University of Chicago press, Chicago.

[2] Clavius, C. 1992. *Corrispondenza*. Baldini, U. and Napolitani, P.D., Eds. Università di Pisa, Dipartimento di matematica, Pisa.

[3] Barreno, P.G. 1995. *La Real Academia de Ciencias : 1582-1995*. Real Academia de Ciencias Exactas, Físicas y Naturales, Madrid.

[4] Manning, C.D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

[5] Jurafsky, D. and Martin, J. H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

[6] Smith, D. N. and C.W. Blackwell. 2012. Four URLs, Limitless Apps: Separation of Concerns in the Homer Multitext Architecture. In *Donum natalicium digitaliter confectum Gregorio Nagy septuagenario a discipulis collegis familiaribus oblatum: A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by His Students, Colleagues, and Friends*. The Center of Hellenic Studies of Harvard University.

[7] Bozzi, A. and Cappelli, G. 1990. A project for Latin Lexicography: 2. A Latin morphological analyzer. *Computers and the Humanities* 24, 5-6, 421-426.

[8] Crane, G. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing* 6, 4, 243-245.

[9] Halácsy, P., Kornai, A., and Oravecz, C. 2007. HunPos: An Open Source Trigram Tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, 209-212, Stroudsburg, PA, USA. Association for Computational Linguistics.

[10] Bamman, D. and Crane, G. 2008. Building a Dynamic Lexicon from a Digital Library. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, 11-20. ACM, New York, NY, USA.

[11] Bamman, D. and Crane, G. 2011. The Ancient Greek and Latin Dependency Treebanks. In *Language Technology for Cultural Heritage, Theory and Applications of Natural Language Processing*, C. Sporleder, A. Bosch and K. Zervanou, Eds.Springer Berlin, Heidelberg, 79-98.

[12] Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., and Zampolli, A. 2000. SIMPLE: A General Framework for the development of Multilingual Lexicons. *International Journal of Lexicography*, special issue, Dictionaries, 13, 4, 249-263.

[13] Pustejovsky, J. 1995. *The Generative Lexicon*. The MIT Press, Cambridge MA.

[14] Treisman, A. 1985. Preattentive processing in vision. *Computer vision, graphics, and image processing* 31, 2, 156-177.

[15] Camarda, D. V., Mazzini, S., and Antonuccio, A. 2012. Lodlive, exploring the web of data. In *Proceedings of the 8th International Conference on Semantic Systems*. ACM, 197-200.