Enriching Digitized Medieval Manuscripts: Linking Image, Text and Lexical Knowledge

Aitor Arronte Álvarez
Center for Language and
Technology
University of Hawaii
arronte@hawaii.edu

Abstract

This paper describes an on-going project of transcribing and annotating digitized manuscripts of medieval Spanish with paleographic and lexical information. We link lexical units from the manuscripts with the Multilingual Central Repository (MCR), making terms retrievable by any of the languages that integrate MCR. The goal of the project is twofold: creating a paleographic knowledge base from digitized medieval facsimiles, that will allow paleographers, philologist, historical linguist, and humanities scholars in general, to analyze and retrieve graphemic, lexical and textual information from historical documents; and on the other hand, developing machine readable documents that will link image representations of graphemic and lexical units in a facsimile with Linked Open Data resources. This paper concentrates on the encoding and cross-linking procedures.

1 Introduction

In recent years, historical documents have been massively digitized and published online in openly available databases, gathering much of the attention of the Digital Humanities community. As a result, large collections of historical handwriting online databases have emerged such as Pares[1], paleographic resources like DigiPal[2], citizen scholar projects (Deciphering Secrets: Unlocking the Manuscripts of Medieval Spain[3]) and digital paleography learning tools (Spanish Paleography Digital Teaching and Learning Tool[4]). In this context, computerized tools have become part of the toolkit of the current humanities scholar.

Most of the research in the computational analysis of digitized historical handwritten documents, has concentrated in its paleographic analysis: the deciphering, dating, and description of ancient manuscripts (Wolf, et. al, 2011; Hassner, et. al, 2013). In this paper, we describe an ongoing project for encoding digitized medieval Spanish manuscripts from the 13th, 14th and early 15th centuries, and linking their content with the Multilingual Central Repository (MCR)[5] (Gonzalez-Agirre et al., 2012).

The main goal of the project is the development of an online database of digitized medieval manuscripts that will enable users to obtain graphemic and lexical information from facsimiles. Manuscripts will be fully searchable using any of the languages that integrate the MCR.

The resource will aid the paleographic understanding of medieval manuscripts as well as the linguistic and philological analysis of medieval Spanish. Also, the database can be a valuable source for computational researchers interested in the automatic processing of medieval manuscripts, since image data will be linked to text and lexical information. To our knowledge, an online resource of this type does not exist.

In this paper we concentrate on: the description of the methods for transcribing, annotating, and encoding manuscripts; the process of automatically linking their content at a lexical level with MCR entries, and for codifying these relationships in a model.

2 Encoding transcriptions of medieval manuscripts

Historical Spanish language varieties exhibit important differences not only at the syntactical and

[1] http://pares.mcu.es/

[2] http://www.digipal.eu/

[3] http://decipheringsecrets.net/

[4] http://spanishpaleographytool.org/

[5] http://adimen.si.ehu.es/web/MCR/

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 73–77,
Beijing, China, July 30, 2015.

morphological level, but also at the graphemic. This is due to the fact that orthographic rules in Spanish were not defined until the 18th century[6], bringing serious difficulties for the understanding of Medieval Spanish[7] manuscripts, since there is substantial variation even within documents of the same period; mostly because scribes had different handwriting styles. Medieval orthography also does not follow contemporary patterns, there is not in a strict sense, different options between graphemes, but rather a combination of factors that may explain certain solutions. As mentioned by Sánchez-Prieto (2004), medieval manuscripts should be understood following a triple correlation of factors:

1. Paleographic uses and shapes of the letters.
2. Identification of the letters.
3. Phonetic changes.

In this triple relation lies the evolution of handwriting, and may reveal important aspects of phonetic change. For that reason, handwritten medieval documents are nowadays manually transcribed with the aid of computational tools. In Figure 1, two examples of different handwriting styles from early 15th century are shown, where the grapheme "a" at the end of each word, is written in a triangular shape (for the first word "la"), and in a square shape (for the second word "buena"). In our work, we segmented the transcribed words in manuscripts using the UVic Image Markup Tool[8], which allows for the annotation and transcription of facsimiles using the Text Encoding Initiative (TEI[9]) model. We customize the markup of the TEI document to be able to codify graphemic and lexical information.

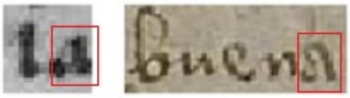

Figure 1: Early 15th century handwriting styles. The grapheme "a" is marked in a box.

TEI is de facto XML standard for the representation of texts in digital form. Following TEI guidelines, different graphemic representations are declared using the element <glyph> in the header of the document (see Figure 2). Image representations of words in the facsimile are segmented using the element <surface>, which defines a written surface as a two-dimensional coordinate space, specifying zones of interest or grouping graphic representations within that space; and the <zone> element, that defines a two dimensional area within a <surface>. The attributes @ulx, @uly, @lrx and @lry, represent respectively, the *x* and *y* values for the upper left and lower right corners of a rectangular space (see Figure 3). Declarations of graphemes are linked to the transcribed text using the element <g>, so variations of a grapheme can be identified and compared. Transcribed words are represented in the TEI document using the <w> element. We automatically generate unique xml:id for each element in the TEI document. The nested representation of words and graphemes in a facsimile is shown in Figure 4.

```
<encodingDesc>
    <charDecl>
      <glyph xml:id="a1">
       <glyphName> roman letter a with
       triangular shape</glyphName>
      <charProp>
        <locaName> entity</localName>
        <value>a1</value>
      </charProp>
      <figure>
        <graphic url="a1.png"/>
      </figure>
      </glyph>
    </charDecl>
 </encodingDesc>
```

Figure 2: Grapheme declaration in TEI document.

```
<facsimile xml:id="imtAnnotatedImage">
   <surface>
     <graphic url="DiegoHernandez.jpg"
width="902px" height="1240px"></graphic>
       <zone xml:id="imtArea_1"
ulx="298" uly="233" lrx="326" lry="253"
rend="visible"></zone>
        <zone xml:id="imtArea_2"
ulx="326" uly="234" lrx="343" lry="251"
rend="visible"></zone>
        <zone xml:id="imtArea_3"
ulx="344" uly="237" lrx="393" lry="254"
rend="visible"></zone>
         <zone xml:id="imtArea_4"
ulx="345" uly="233" lrx="391" lry="252"
rend="visible"></zone>
         <zone xml:id="imtArea_5"
ulx="363" uly="238" lrx="372" lry="251"
rend="visible"></zone>
```

[6] The Royal Spanish Academy published the first orthographic rules in 1726, as part of the first volume of the Diccionario de Autoridades.

[7] Medieval Spanish is an early form of Spanish that ranges from the 10th to the 15th century, as it is considered by most scholars.

[8] http://tapor.uvic.ca/~mholmes/image_markup/index.php

[9] http://www.tei-c.org/index.xml. TEI is an XML-based file format for exchanging text.

```
  </surface>
</facsimile>
```

Figure 3: TEI representation of image segments in a facsimile.

```
<body>
  <div xml:id="imtImageAnnotations">
    <s xml:lang="spa">
     <w xml:id="ms1_w_6" cor-
resp="#imtArea_1"> por</w>
      <w xml:id="ms1_w_7" cor-
resp="#imtArea_2">la</w>
      <w xml:id="ms1_w_8" cor-
resp="#imtArea_3">gr<g xml:id="ms1_g_1"
corresp="#imtArea_4"
ref="#a1">a</g>çia</w>
       <w xml:id="ms1_w_9" cor-
resp="#imtArea_5">de</w>
     </s>
  </div>
</body>
```

Figure 4: text representation of words and graphemes linked to their corresponding image segments.

3 Linking medieval manuscripts with multilingual lexical resources

In order to link image representations of words in a historical variety with a multilingual lexical database, two operations need to take place:

1. Matching the historical form of the word with its contemporary standard.
2. Codifying that relation in the document.

3.1 Mapping medieval Spanish with contemporary standard

Before the cross-linking of the transcribed words from the manuscript with MCR entries, words from medieval Spanish will need to be mapped to the standard form of contemporary Spanish. We follow the rules presented in (Sánchez-Prieto, 2004) and previous computational work on historical language varieties (Sánchez-Marco et. al, 2010).

The mapping rules used can be divided into substring rules and word rules. In Table 1 examples of the mappings using substring rules are introduced. Words that are not covered by the substring rules include graphemic variation of the type: *decaydo→decaído, fablar→hablar.*

Substring rules		
Medieval	Modern	Transformation
euo	*evo*	*nueuo→nuevo*
iua	*iva*	*dadiua→dadiva*

Table 1: substring rules

3.2 Linking terms from medieval manuscripts with MCR synsets

WordNet is a large lexical database of English (Miller, 1995), where open class words are grouped into concepts represented by synonyms (synsets) that are linked to each other by semantic relations such as hyponymy and meronymy. There are multiple wordnets for different languages, and wordnets for groups of languages like the Euro WordNet (Vossen, 1998). Wordnets have also been extended by using external lexical resources like Wiktionary (McCrae et al., 2012) or with a combination of multilingual resources (De Melo & Weikum, 2009; Bond & Foster, 2013). Also, the Portuguese wordnet incorporates non standard varieties of the language (Marrafa et al., 2011). Our goal is to link words from historical varieties of Spanish extracted from manuscripts, to synsets in MCR, in such a way that the image representation of a medieval word can be directly associated to its contemporary form or via semantic relations to a sense.

The MCR integrates wordnets in five different languages, English, Spanish, Catalan, Basque and Galician that are linked to each other via an Inter-Lingual-Index (ILI). Each of the wordnets in the MCR is aligned to the Princeton WordNet 3.0 and encoded using Lexicon Model for Ontologies (*lemon*)[10]. One of the main advantages of using *lemon* is its linguistically sound structure based on the Lexical Markup Framework (LMF), making it an ideal model for lexicons and machine-readable dictionaries in the Linked Data Cloud.

We use a Python script for linking words from historical Spanish manuscripts encoded in a TEI document with existing MCR synsets, by matching lemmas in medieval Spanish with contemporary standard. We follow the substring and word mapping rules described in the previous subsection, matching them with contemporary Spanish lemmas in MCR. This approach is imperfect, since there are medieval words that no longer exist or might have different lexical realizations. In these cases, medieval words will need to be

[10] http://lemon-model.net/

linked with synsets via the linguistic analysis of their meanings using a historical dictionary, following an approach similar to the one described in (Declerck et al., 2014). Also, we should note that PoS-taggers of standard contemporary Spanish used in a historical variety context, perform below state of the art taggers (Sánchez-Marco et. al, 2011), which makes manual verification an unavoidable step.

In order to represent the semantic linking of the words in the TEI document with lexical entries in the *lemon* model of the MCR, we need to extend our initial TEI representation with pointing mechanisms associated to the TEI <relation> element. In Figure 5 we show how semantic relationships can be established between words in the manuscripts and external lexical resources.

```
<relation
 ref="http://www.lemon-
 model.net/lemon#formVariant"
 active="#ms1_w_8"
passive =" mcr:spa-gracia-n#Sense-
04840715-n "/>

<relation
 ref="http://wordnet-
rdf.princeton.edu/ontology#translation"
 active="http://wordnet-
rdf.princeton.edu/wn30/14458181-n"
 passive="#ms1_w_8"/>
```

Figure 5: semantic annotations to external resources.

4 Next steps: sharing knowledge between manuscripts

In the encoding presented in this paper, manuscripts are annotated, codified, and linked to external lexical resources. Even though several paleographic and graphemic relations are established implicitly in the markup of the TEI document, this representation of a manuscript does not provide semantic relationships beyond the ones defined at a lexical level. In order to share paleographic knowledge with other open resources across the web, following Linked Open Data principles (Chiarcos et. al, 2011), a paleographic ontology for medieval Spanish documents needs to be develop. The ontology should capture relationships within a given document, between different manuscripts in a collection and between different collections. Allographs, glyphs, ligatures, word and common name abbreviations, contractions, acronyms, numbers and dates variations in notation should be defined at an ontological level. More general relationships and document data such as typology (legal, church, private document, etc), style, place of origin, manuscript collection, archive, author and year of the transcription, will also be included in the ontology.

Since we are dealing with cultural heritage materials, existing ontologies such as the Functional Requirements for Bibliographic Records (FRBR), CIDOC Conceptual Reference Model (CIDOC CRM), and more directly related to our work, the General Ontology for Linguistic Description (GOLD), already define the terminology and some of the relationships that can be found in medieval manuscripts; in these cases classes can be directly imported and reused. In some other cases, classes may need to be created to define specific graphemic objects and paleographic relationships that are not defined in the existing ontologies.

In the TEI representation described in this paper, unique ids are given to lexical and graphemic units, allowing for the automatic creation of URIs that can be used for external resources to link to it. The RDF annotation can be done following the example given in Figure 5 using the element <relation> and the relationships defined in the paleographic ontology.

At a lexical level, semantic relations are established in the TEI document via external resources. Even though lexical ontologies such as *lemon,* and to some extent WordNet, define linguistic relationships between lexical units, they may not be specific enough to describe the different relationships of a language with its historical varieties. Future steps in the project should consider adding such detailed linguistic relationships.

5 Conclusions

In this paper we described the first steps towards the creation of an online resource of digitized medieval Spanish manuscripts, where graphemic, lexical and textual information can be retrieved directly from facsimiles. We have shown and demonstrated a method for transcribing and encoding in TEI P5 image data from manuscripts. We have described also how medieval Spanish can be linked to its contemporary standard and to the rest of the languages that integrate MCR, making manuscript terms retrievable using any of these languages. Next steps in the project include: developing a paleographic ontology of medieval Spanish, extending semantic annotations at a lexical level incorporating historical varieties relationships, building a web interface, and making data available in the cloud.

References

Bond, F., & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. *ACL, 1*, pp. 1352-1362.

Chiarcos, C., Hellmann, S., & Nordhoff, S. (2011). Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. *52* (3), 245-275.

De Melo, G., & Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 513-522). ACM.

Declerck, T., Wand-Vogt, E., Mörth, K., & Resch, C. (2014). Towards a Unified Approach for Publishing Regional and Historical Language Resources on the Linked Data Framework. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data, 17.*

Gonzalez-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual Central Repository version 3.0. *LREC*, (pp. 2525-2529).

Hassner, T., Rehbein, M., Stokes, P., & Wolf, L. (2013). Computation and Palaeography: Potentials and Limits. *Dagstuhl Manifestos , 2* (1), 14-35.

Marrafa, P., Amaro, P., & Mendes, S. (2011). WordNet. PT global: extending WordNet. PT to Portuguese varieties. *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties* (pp. 70-74). Association for Computational Linguistics.

McCrae, J., Montiel-Ponsoda, E., & Cimiano, P. (2012). Integrating WordNet and Wiktionary with lemon. In *Linked Data in Linguistics* (pp. 25-34). Berlin, Heidelberg: Springer.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM , 38* (11), 39-41.

Sánchez-Marco, C., Fontana, J., Domingo, J., & Boleda Torrent, G. (2010). *Annotation and representation of a diachronic corpus of Spanish.*

Sánchez-Marco, C., Boleda, G., & Padró, L. (2011). Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 1-9). Association for Computational Linguistics.

Sánchez-Prieto, P. (2004). *La normalización del castellano escrito en el siglo XIII. Los caracteres de la lengua: grafías y fonemas.*

Vossen, P. (1998). *A multilingual database with lexical semantic networks.* Dordrecht: Kluwer Academic Publishers.

Wolf, L., Potikha, L., Dershowitz, N., Shweka, R., & Choueka, Y. (2011). Computerized paleography: tools for historical manuscripts. *18th IEEE International Conference on Image Processing (ICIP)* (pp. 3545-3548). IEEE.