Editing a XVth century political treatise using the computer: a back-and-forth between meaning and information

Matthias Gille Levenson
PhD student, École Normale Supérieure de Lyon

Iberian Connections seminar

November 12, 2019
Information and meaning

Ms. 2097, University of Salamanca
fol. 436r

Inc/901, National Library, Madrid
fol. 244v

Ms. II/215, Real Biblioteca, Madrid
fol. 453r

4 contar JRZG | notar Q 4 muy JQZG | om. R 5 muy JZ | om. RQG 5 judios JRQG | indios Z 5 era de Jherusalém J | era de Iherusalem R era de Iherusalém Q era gran pintor Z era de Gerusalén G 5–6 las ystorias de JR | las ystorias de Q las ystorias de Z om. G 6–7 E fizo que pintase sobre la sepultura de la dicha Reyna
Acquiring the information: the transcription. To OCR (HTR?) or not to OCR

Advantages:
- Gain of time for large corpuses
- Conservation of graphical features made easier

Method:
1. Make a conservative transcription of some folios of the witness;
2. Feed the program with the transcription = train a model with Ocropy [Breuel 2008];
3. Predict new text, correct, re-train, and so on until a given error rate is reached;
4. Use the best model on new folios.

Results:
- Low error rate with incunabulas (5%);
- Less accurate with manuscript writing, but it is improving: Kraken [Kiessling 2019];
- The main issue is the line segmentation.
Acquiring the information: the transcription. To OCR (HTR?) or not to OCR

- **Advantages:**
  - Gain of time for large corpuses
  - Conservation of graphical features made easier

**Method:**
1. Make a conservative transcription of some folios of the witness;
2. Feed the program with the transcription = train a model with Ocropy [Breuel 2008];
3. Predict new text, correct, re-train, and so on until a given error rate is reached;
4. Use the best model on new folios.

**Results:**
- Low error rate with incunabulas (5%);
- Less accurate with manuscript writing, but it is improving: Kraken [Kiessling 2019];
- The main issue is the line segmentation.
Acquiring the information: the transcription. To OCR (HTR?) or not to OCR

• **Advantages:**
  - Gain of time for large corpuses
  - Conservation of graphical features made easier

• **Method:**
  1. Make a conservative transcription of some folios of the witness;
  2. Feed the program with the transcription = train a model with Ocropy [Breuel 2008];
  3. Predict new text, correct, re-train, and so on until a given error rate is reached;
  4. Use the best model on new folios.

Results:

  - Low error rate with incunabulas (5%);
  - Less accurate with manuscript writing, but it is improving: Kraken [Kiessling 2019];
  - The main issue is the line segmentation.
Acquiring the information: the transcription. To OCR (HTR?) or not to OCR

- **Advantages:**
  - Gain of time for large corpuses
  - Conservation of graphical features made easier

- **Method:**
  1. Make a conservative transcription of some folios of the witness;
  2. Feed the program with the transcription = train a model with Ocropy [Breuel 2008];
  3. Predict new text, correct, re-train, and so on until a given error rate is reached;
  4. Use the best model on new folios.

- **Results:**
  - Low error rate with incunabulas (∼ 5%);
  - Less accurate with manuscript writing, but it is improving: Kraken [Kiessling 2019];
  - The main issue is the line segmentation.
Structuring the information: the TEI

Matthias Gilles Levenson

From information to meaning

November 12, 2019 4 / 22
Structuring the information: the TEI

What are the interests of a community driven standard? [Burnard 2015]
What are the interests of a community driven standard? [Burnard 2015]

- It’s a standard!
What are the interests of a community driven standard? [Burnard 2015]

- It’s a standard!
- And it’s community driven.
What are the interests of a community driven standard? [Burnard 2015]

- It’s a standard!
- And it’s community driven.
- An *ontology on the structure of texts*¹, a “conceptual model of textuality” [Ciotti 2018].

---

¹*N.B.: It is not* an informatical ontology! See [Ciotti and Tomasi 2016]
Enriching the information: lemmatisation and POStagging

Take *aver, auer, haver*:
Enriching the information: lemmatisation and POStagging

Take *aver, auer, haver*:

- Three different graphies. Form: *aver | auer | haver*
Enriching the information: lemmatisation and POStagging

Take *aver, aper, haver*:

- Three different graphies. **Form**: *aver | aper | haver*
- Three forms of the verb *haber*. **Lemma**: *haber | haber | haber*
Enriching the information: lemmatisation and POStagging

Take *aver, auer, haver*:

- Three different graphies. **Form**: *aver | auer | haver*
- Three forms of the verb *haber*. **Lemma**: *haber | haber | haber*
- Three infinitives. **Part Of Speech**: *VMN000 | VMN000 | VMN000 [EAGLES / FREELING]*

I'm using the dictionary created by Sánchez Marco for her PhD dissertation [Sánchez Marco 2012].
Enriching the information: lemmatisation and POStagging

Take *aver, auer, haver*:

- Three different graphies. **Form**: *aver | auer | haver*
- Three forms of the verb *haber*. **Lemma**: *haber | haber | haber*
- Three infinitives. **Part Of Speech**: *VMN000 | VMN000 | VMN000 [EAGLES / FREELING]*

<table>
<thead>
<tr>
<th>Form</th>
<th>Lemma</th>
<th>PoS</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>aver, auer, haver</em></td>
<td>HABER</td>
<td>VMN000</td>
</tr>
</tbody>
</table>
Enriching the information: lemmatisation and POStagging

Take *aver, auer, haver*:

- Three different graphies. Form: *aver | auer | haver*
- Three forms of the verb *haber*. Lemma: *haber | haber | haber*
- Three infinitives. Part Of Speech: *VMN000 | VMN000 | VMN000* [EAGLES / FREELING]

\[
\begin{array}{ccc}
\text{FORM} & \Rightarrow & \text{LEXMA} & \text{PoS} \\
\text{aver, auer, haver} & \Rightarrow & \text{haber} & \text{VMN000} \\
\end{array}
\]

This grammatical information is added to the TEI encoding, to be processed after.
Enriching the information: lemmatisation and POS tagging

Take *aver, auer, haver*:

- Three different graphies. **Form**: *aver | auer | haver*
- Three forms of the verb *haber*. **Lemma**: *haber | haber | haber*
- Three infinitives. **Part Of Speech**: *VMN000 | VMN000 | VMN000* [EAGLES / FREELING]

<table>
<thead>
<tr>
<th><strong>Form</strong></th>
<th><strong>Lemma</strong></th>
<th><strong>PoS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>aver, auer, haver</em></td>
<td><em>haber</em></td>
<td><em>VMN000</em></td>
</tr>
</tbody>
</table>

This grammatical information is added to the TEI encoding, to be processed after.

```
<w lemma="haber" pos="VMN000">aver</w>
<w lemma="caballero" pos="NCMP000">cavalleros</w>
<w lemma="muy" pos="RG">muy</w>
```

I’m using the dictionary created by Sánchez Marco for her PhD dissertation [Sánchez Marco 2012].
What is the *collatio*?

“La colación o cotejo de todos los testimonios entre sí para determinar las lectiones variae o variantes”.

[Blcua 1983]

Can we simulate it with a computer? Let’s highlight the two steps of the *collatio*:
What is the *collatio*?

“La colación o cotejo de todos los testimonios entre sí para determinar las lectiones variae o variantes”.

[Blecua 1983]

Can we simulate it with a computer? Let’s highlight the two steps of the *collatio*:

1. Finding the portion of text to be compared in each witness
2. Making the comparison
What is the *collatio*?

“La colación o cotejo de todos los testimonios entre sí para determinar las lectiones variae o variantes”.

[Blécua 1983]

Can we simulate it with a computer? Let’s highlight the two steps of the *collatio*:

1. Finding the portion of text to be compared in each witness
2. Making the comparison

The human mind *does not dissociate* these two steps, but the computer needs this distinction.
Comparing it and eliminating the redundancy I: the alignment

Alignment (= search for similar groups of words) on the forms with CollateX [Dekker and Middell 2011]

1. “quéales e quántas cosas deuen auer los buenos lidiadores”: base sentence
2. “quéales e quántas cosas deven aver los buenos lidiadores”: 2 differences
3. “quéales e quántas cosas deven haver los buenos lidiadores”: 2 differences
Comparing it and eliminating the redundancy I: the alignment

Alignment (= search for similar groups of words) on the forms with CollateX [Dekker and Middell 2011]

1. “quáles e quántas cosas deuen auer los buenos lidiadores”: base sentence
2. “quáles e quántas cosas deven aver los buenos lidiadores”: 2 differences
3. “quáles e quántas cosas deven haver los buenos lidiadores”: 2 differences

Result:

<table>
<thead>
<tr>
<th>quáles e quántas cosas</th>
<th>deuen auer</th>
<th>omisit</th>
<th>los buenos lidiadores</th>
</tr>
</thead>
<tbody>
<tr>
<td>quáles e quántas cosas</td>
<td>deven</td>
<td>aver</td>
<td>los buenos lidiadores</td>
</tr>
<tr>
<td>quáles e quántas cosas</td>
<td>deven</td>
<td>haver</td>
<td>los buenos lidiadores</td>
</tr>
</tbody>
</table>
Comparing it and eliminating the redundancy I: the alignment

Alignment on the lemmas with CollateX

1. cual + y + quanto + cosa + deber + haber + el + buen + lidiador : base sentence represented as lemmas
2. cual + y + quanto + cosa + deber + haber + el + buen + lidiador: no difference
3. cual + y + quanto + cosa + deber + haber + el + buen + lidiador: no difference

Result:

| quéales e quéantas cosas | deuen | auer los buenos lidiadores p |
| quéales e quéantas cosas | deven | aver los buenos lidiadores p |
| quéales e quéantas cosas | deven | haver los buenos lidiadores |
Comparing aligned groups of words: is there a variation?

For each aligned group:

1. If the strings (= the characters) are the same, it is not a variant: no apparatus entry
2. If the strings are different, we have a variant.

We are talking about strings here, not about words! It is pure information. Can we go further? What can we do with the variants?
Comparing it and eliminating the redundancy II: the comparison

Improving the accuracy of the apparatus: graphical variants identification

<table>
<thead>
<tr>
<th>Form</th>
<th>Lemma</th>
<th>PoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>aver, auer, haver</td>
<td>HABER</td>
<td>VMN000</td>
</tr>
</tbody>
</table>
Comparing it and eliminating the redundancy II: the comparison

Improving the accuracy of the apparatus: graphical variants identification

<table>
<thead>
<tr>
<th>FORM</th>
<th>( \Rightarrow )</th>
<th>LEMMA</th>
<th>PoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>aver, auer, haver</td>
<td>( \Rightarrow )</td>
<td>haber</td>
<td>VMN000</td>
</tr>
</tbody>
</table>

**Meaning**

“Aver, auer, haver are the same word…”

“When you have a graphical variant, do this” (Method)

**Information**

“These three tokens have different strings, the same lemma and the same POS”

“If he strings are not equal AND their lemma is the same AND so is the PoS: do this” (Algorithm)
To sum up

1. Align...
   1.1 Alignment on the lemmas

2. ... And compare. **Algorithm:** for each aligned token or group of token:
   2.1 if all strings are strictly equal, we haven’t got a variant.
   2.2 if the strings are different, it is a variant.
      But this is not enough:
         2.2.1 if the words have the same lemma and the same POS, we have a graphical variant ! (\(> 25\%\))
         2.2.2 if the lemma (or the POS) differ, we have a “real” variation.

The result of the process will be encoded in TEI, and will be injected to the individual transcriptions.
Going back and forth

Figure 1. Human-readable, consistent, standard information
" Going back and forth

Figure 2. Human-unreadable information
Figure 3. Human-readable, consistent, standard information
Translating the information: the output document. *The meaning?*

Transformation of the XML into \LaTeX\ or to a web-based interface.
Mas aquí podemos contar de cómo Alexandre fizo muy grant onrra a la muger de Darío. E llamó y un muy grant sabio de los judíos que dezían Apelles, que era de Jherusalém, e sabía muy bien las ystorias de la bibla. E fizo que pintase sobre la sepultura de la dicha Reyna todos los fechos granados.

4 contar JRZG | notar Q  4 muy JZG | om. R  5 muy JZ | om. RQG  5 judios JRQG | indios Z  5 era de Jherusalém, J | era de Iherusalém R era de Iherusalém Q era gran pintor Z era de Gerusalén G  6 E J | om. RQ e Z et G  6-7 fizo que pintase sobre la sepultura de la dicha Reyna todos los fechos granados que JG | om. RQ fizo que pintasse
Results

Ms. 2097, University of Salamanca fol. 436r

Ms. II/215, Real Biblioteca, Madrid fol. 453r

Inc/901, National Library, Madrid fol. 244v

Mas aquí podemos contar de cómo Alexandre fizo muy gran honra a la mujer de Darío: et llamó y un grand sabio de los judíos que dezían Apelles, que era de Gerusalén, e sabía muy bien las estorias de la Blivia; et fizo que pintase sobre la sepultura de la dicha Reyna todos los fechos.

4 contar JRZG | notar Q 4 muy JQZG | om. R 5 judios JRGQ | indios Z 5 era de Gerusalén, G | era de Jherusalém J era de Iherusalem R era de Iherusalém Q era gran pintor Z 6 et G | E J om. RQ e Z 6–7 fizo que pintase sobre la sepultura de la dicha Reyna todos los fechos granados que JG | om. RQ fizo que pintasse sobre la sepultura de la
Mas aquí podemos contar de como Alixandre fizo muy grand honra a la muger de Dario: ĵ llamó y un muy grand sabio de los indios que dezian Apelles que era gran pintor: ĵ sabía muy bien las hystorias de la brivia: ĵ fizo que pintasse sobre la sepultura de la dicha reyna: todos los fechos
Conclusions

Correction of the text, silenciation of the Jewish heritage, or a bit of both?
What comes next?

Since we cannot avoid considering the text as information...

- **Accessibility**: DTS, a IIIF-like standard API for texts.
- **Citability**: What to do with the revisions of a digital work?
- **Identification of passages**: When we cite a passage, do we have to cite the page or its identifier?
- **Perennity**: web-based interfaces are really hard to maintain over the time [Pierazzo 2015, pp. 173–179]
Bibliography


