Semantic Austen

Jane Austen Semantic Digital Library and annotation campaign

Nicol D'Amelio - nicol.damelio@studio.unibo.it

Maria Juliana Gamboa Nivia - mariajuliana.gamboa@studio.unibo.it

Maryam Dadrasrazi – maryam.dadrasrazi@studio.unibo.it

Annotation Campaign

Annotation Pipeline

Task Definition

We performed an annotation campaign on Jane Austen's manuscripts from different moments in her life to create a homogeneous batch of texts that cover her lifetime as an author. For the content of the manuscripts, we took the diplomatic transcriptions done in the Scholarly Digital Edition (SDE) *Jane Austen Fiction Manuscripts*[1].

Our aim was to create a Model that could generate an automatic transcription of all of Jane Austen's manuscripts in our potential Semantic Digital Library *Semantic Austen* that includes the work done in the SDE. Therefore, we created this model on Austen's handwriting so that any other existing holograph text of her can be transcribed using it.

Context

Jane Austen (1775-1817) was a 19th century English writer; her manuscripts are in possession of several institutions that digitize them for online public visualization.

We chose to work on three manuscripts available online:

- *Juvenilia*: (1787-1793), Bodleian Library, Oxford
- *The Watsons:* (1803), Bodleian Library, Oxford
- *Persuasion*: (1817), British Library

The images of the Bodleian Library[2] are under the CC-BY-NC license and those from the British Library[3] are of public domain.

The SDE *Jane Austen Fiction Manuscripts* made an extensive philological work providing the diplomatic transcriptions of the texts, that we follow for the annotation of the manuscripts.

Although the SDE made extensive work, it is slightly outdated and doesn't make use of semantic and interoperable elements. We would like to create a Semantic Digital Library called *Semantic Austen* based on the work of the SDE and take it forward, providing images with IIIF, texts to be downloadable in XML, external links related to Jane Austen.

---

[1] Jane Austen fiction Manuscripts: https://janeausten.ac.uk/index.html.
[2] Bodleian Library, Oxford: rights https://digital.bodleian.ox.ac.uk/terms/
[3] British Library: rights https://www.britishlibrary.cn/en/works/jane-austens-juvenilia/

Corpus selection

Juvenilia, The Watsons, Persuasion are representative of different periods of Austen's work and handwriting (from childhood to maturity) and were selected to train the model on different ways of her handwriting.

We began with a 50-page sample but expanded it to improve model performance to 68 annotated pages (28 from *Juvenilia*, 25 from *The Watsons* and 15 from *Persuasion*) with 10522 words.

Annotation guidelines

Our campaign acts in accordance with the OCR and Transkribus guidelines[4].

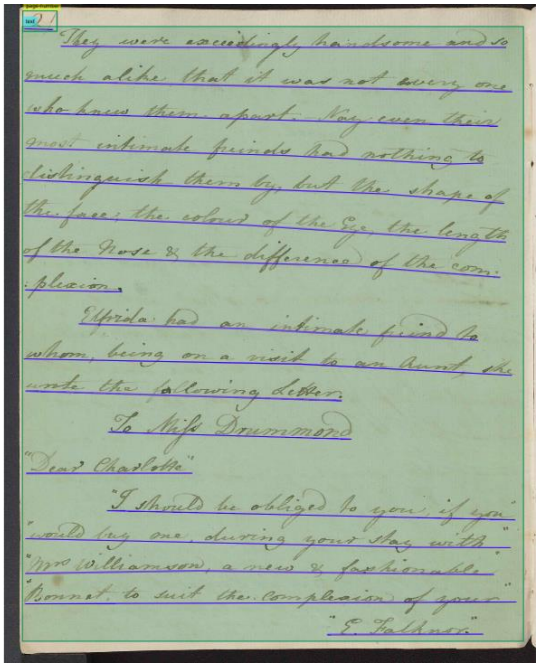
We relied on the diplomatic transcription of the SDE for the transcription of the texts. This aligns with level 2 of the OCR-D Guidelines for Ground Truth[5], to preserve the original text as accurately as possible.

We defined some internal guidelines for special cases often encountered.

Layout guidelines

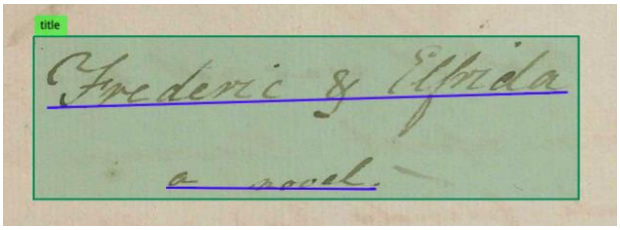
We defined custom made regions to distinguish structural elements of the page:

1. Text:

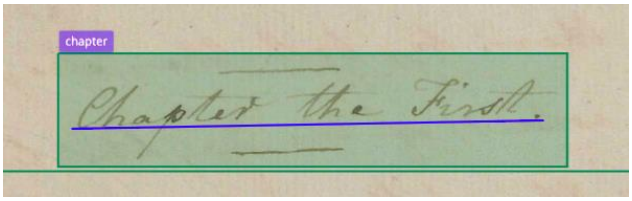


[4] Transkribus guidelines: https://help.transkribus.org/
[5] OCR-D Guidelines for Ground Truth: https://ocr-d.de/en/gt-guidelines/trans/level_2_2.html

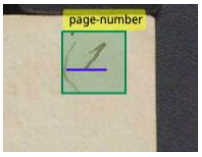
2. Title:



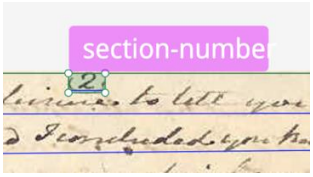
3. Chapter:



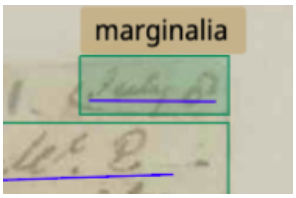
4. Page-number:



5. Section number: numbered quires



6. Marginalia: where she includes dates



7. All lines traced on the page that are not a baseline of text were ignored.

Text guidelines

The handwritten text conveyed particularities that had to be specified manually:

1. Corrections on top of erased text were defined in an extra baseline; the erased text was tagged as *strikethrough* while we tagged the correct text on top as *superscript;*

2. Additions of text (^) were defined in an extra baseline and tagged as *superscript* and described with the custom tag *addition;*
3. In cases of erased text on top of which is a correction also pointed out as addition by (^), we only tag the correction as *superscript;*
4. We didn't include (^) but expressed it with the *addition* tag;
5. Letters written with pencil and later erased weren't transcribed;
6. We follow diplomatic transcription in punctuation;
7. Normalization of words; e.g. M^{r} to Mr.

Pilot:

The pilot was performed to understand how efficient our model was, since we didn't use a pre-existing English HTR model. Using 50 pages between the three novels made us understand we had to make the dataset bigger to receive the results that we expected.

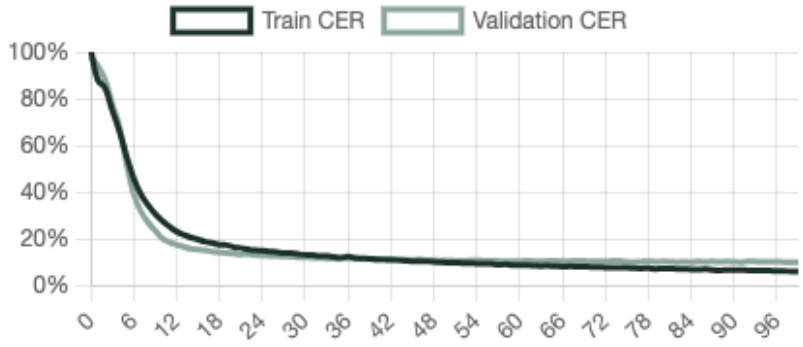
Campaign:

The campaign was conducted using Transkribus: we annotated the pilot (50 pages) with the transcriptions of the SDE and then trained our first model selecting 10% as the validation set obtaining a Character Error Rate (CER) of 17.03%.

We made a re-training of our model with our pilot adding a small-set of annotated pages expanding the corpus to 68 pages, obtaining a CER of 16,47%.

We performed an additional fine-tuning with "Transkribus English Handwriting M3" base model, however the CER increased to 18% due to overfitting.

We did one last modification to the parameters, increasing the validation set from 10% to 16%, while maintaining 100 epochs and we finally reached a CER of 10,11%.



Evaluation of Results:

A CER of 10.11%, is considered highly functional, since results of 10% or under are considered very efficient. It produces reliable base text that requires minor manual intervention.

Comparison:

To understand the efficiency of our HTR model we tested Gemini 3 flash OCR, by providing the images without any context.

To compare the results of Gemini against what our HTR model could do, we provided both with 6 new pages (not used for the training model), 2 from each novel.

Gemini gave exceptional results, since it was able to understand the provenance of the text and identified the specific part of the novel. It provided highly accurate if not perfect text recognition, handling layout, cancellations and superscript corrections.

Transkribus was very accurate, though it still struggled with the constant cancellations and corrections and in one page it didn't understand where the text began on a page missing part of the text.

Conclusions of Annotation Campaign:

Ultimately the model created was a very good starting point for what concerns an annotation campaign dealing with difficult corrections and layout. However, we realized that an LLM such as Gemini trained with thousands of texts has great performances and almost perfect results.

Therefore, for a Semantic Digital Library the use of an LLM with OCR would get better results than our model, yet this annotation campaign is just a starting point for what can be done in the future and for the work we could create for our SDL *Semantic Austen*.

Publication and use

In our annotation campaign, we tried to be compliant with the FAIR Principles.

Semantic Digital Library

As mentioned before, our aim was to create the Semantic Digital Library *Semantic Austen* built upon the work done by the SDE, taking it forward by incorporating semantic elements.

As a sample of how we would like to make it semantic, we created a prototype SDL.

Ideally, the pages included in the collection can be navigated with full-text searches or through a faceted research that reuses the XML tags, to be transformed into RDF triples, describing the sex of the fictional characters, the places and the themes encountered in the novels.

As for the visualization of the single item, we implemented an example using four pages from *Persuasion*. The XML/TEI encoding is used for the HTML rendering of the transcription (through XSLT) alongside the manuscript's image but it is also made available to the user for download. This was not allowed by the SDE, which declared the edition to be XML/TEI encoded but didn't provide access to it.

The encoding was indeed performed by exporting the annotated pages from Transkribus in XML/TEI format and enriching the <teiHeader> of all the unique IDs assigned to all characters and places encountered in the pages. A taxonomy was also created to describe the most recurrent themes in Austen's work: Economics & Property, Ethics & Sentiment, Social Dynamics & Reputation, Space & Communication. Words and sentences were assigned to each of these categories, allowing the users to potentially visualize single pages related to the themes.

The XML to RDF transformation would have to be automatically produced. In our prototype, we created a small Knowledge Graph (.ttl) to be potentially expanded, able to represent all the entities pointed out by the XML tags and the different FRBR levels of the work as presented in the SDL. The serialization reuses properties from the most common Semantic Web ontologies (i.e. DCTerms, schema.org) and includes links to authority files for all entities (VIAF, Wikidata, Geonames). The metadata are also RDFa compliant and reference the same entities described in the XML tags and Knowledge Graph with the same URIs.

We also included a potential IIIF connection in the prototype: since a native manifest from the British Library was not available for these pages, we simulated the connection to show interoperability.

Moreover, we thought of enhancing the user experience of the texts by integrating related media of the novels, referencing linked movies, theatre productions or exhibitions. This feature could ideally be AI-generated, to provide deeper insights on the novel or external contextual information.

Semantic Austen is therefore a project for an annotated Semantic Digital Library, integrating HTR model training and Semantic Web technologies. A potential future expansion could contribute to making it a great state of the art project.

Sitography

- Bodleian Library MS. Don. e. 7: https://digital.bodleian.ox.ac.uk/objects/9e416479-53f1-480f-ac97-de44dd5e7e59/surfaces/23ae6f76-1b4a-4462-b769-63d704f6b0f9/
- British Library Jane Austen: https://events.bl.uk/exhibitions/jane-austen-at-250
- DCTerms: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
- Geonames: https://www.geonames.org
- Jane Austen Fiction Manuscripts: https://janeausten.ac.uk/index.html
- OCR-D Guidelines for Ground Truth: https://ocr-d.de/en/gt-guidelines/trans/level_2_2.html
- schema.org: https://schema.org
- TEI guidelines: https://tei-c.org/guidelines/
- Tranksribus Lite : https://help.transkribus.org/
- VIAF: https://viaf.org
- Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page