Workflow Proposal

Cost Action CA16204 – WG1 *2017-11-04*

Workflow

The objective of this document is to set out the main steps to build the ELTeC core corpus of novels.

The corpus will include novels written during 1850 and 1920 in these languages : Dutch, English, French, German, Modern Greek, Italian, Polish, Portuguese, Russian and Spanish. Each novel will be in a machine readable format and encoded in XML (following standard TEI).

The main steps to achieve this objective are the next:

1. Selecting authors and novels.
2. Finding the novels.
3. Cleaning and normalizing texts.
4. Annotation.
5. Publication.
6. Evaluation.

1 Starting point

The starting point are three documents:

Doc1: *Sampling Criteria*, in which the main requirements for text selection are established.

Doc2: *Encoding Guideline*, in which the encoding scheme is defined.

Doc3: *Workflow*, this document.

2 Step 1. Selecting authors and novels.

The objective of this step is to find appropriate titles for each language published during 1850-1920 according to the sampling criteria (Doc1). It will be done in two sub-steps. The first one is to extract, as far as possible, a list of novels published during that period and the amount of reprints. This information could be extracted from different sources as OCLC WorldCat or specific National Libraries of each Country. See Appendix A for a list of catalogs. Each country will complete this list of catalogs with specific sources.

The list of novels will be published in a web page, including information about author's name, title, date and place of the first edition, number of reprints during the period, source from which this information has been extracted, size, topic, etc. See eltecSheet document. At this moment, the information that must be stored is not fixed. We expect suggestions from the whole group.

The second sub-step is the selection of appropriate novels for the ELTeC corpus. From this list, each group will select appropriate novels according to the criteria of Doc1.

For this step, both the creation of the candidate list and the selection of the final novels, we expect the advice of scholars and expert of each literary tradition.

3 Step 2. Finding the novels.

The objective of this step is to obtain the novels selected in the previous step in machine-readable format (as plain text). Applying selection criteria from Doc1 it is possible that not all novels will be in a machine-readable format or even digitized. In order to transform all texts to this format, we suggest to follow the next steps.

First, to look for novels in digital repositories. See Appendix 2 for some of them. In this case, however, it is important that the text retrieved follows the sampling criteria of DOC1. The novel must be in machine-readable format (plain text, HTML, XML, DOC, ODT, RTF, Epub, or similar.), it must be *licensed* under open or free licenses as Creative Commons or similar (according to MoU, the final corpus will be freely available under Creative Commons License) and it must follow the first edition of the novel. See Doc1 for all the sampling criteria.

If it is not possible to find a novel in a machine-readable format, the second option is to look for it in digital libraries, traying to find a digitized version (PDF format, JPG, or similar) that follows the sampling criteria of Doc1. Finally, if the novel has not been digitized ever, it must be done in this step (if we have resources to do it). We hope that this will be done only for very few novels. In both cases, the text must be transformed to a machine-readable format. See Appendix 3 for a list of tools for digitizing texts. This way we will obtain a machine-readable format (plain text) of each novel. The problem is that it is a complex and time-consuming tasks.

Repository of raw texts. In order to make a back-up of the corpus and the creation process, all novels will be stored in a "raw text repository", a text repository with the texts just like they have been found or digitized.

4 Step 3. Cleaning-up and normalizing texts.

Whether novels have been obtained directly in a machine-readable format or if they have been digitized, it is necessary to check the text in order to fix typos and normalize it. Exactly what elements must be normalized will be decide later.

As far as possible, this task will be done automatically. However, a manual review will be necessary.

Open question. Both fixing and normalizing texts are complex and time-consuming tasks. We have not funding to do them. However, the reliability of the "distant" analysis (WG2) will depend on the quality of these texts. We must decide how they will be done.

5 Step 4. Annotation and encoding.

According to Doc2, three kinds of annotation must be developed: metadata annotation, document structure annotation and linguistic/literary annotation. The last one will not be done for the moment.

5.1 Metadata annotation.

Metadata attributes have been specified in the Encoding Scheme (Doc2). In order to introduce in each text, two tasks must be done.

First, a controlled language must be developed to control the values of each attribute and the coherence of the metadata.

Second, the annotation process of the metadata of each novel: to create a well-formed XML document based on TEI and to introduce header elements and attributes with the appropriate value.

We will try to follow a semi-automatic annotation process here. Metadata could be compiled and stored in a CSV file during previous steps (one file per novel). Then a well-formed XML file could be created automatically (Python script?) taking as an input the plain text of the novel and the corresponding CSV file. At the end, all metadata will be checked (see evaluation step).

5.2 Annotation of the structure of the novel.

Doc2 includes tags for representing the structure of the novel (chapters, paragraphs, etc.). Similarly, we will follow a semi-automatic annotation approach, trying to annotate automatically as much as possible.

However, in both cases, annotation must be manually checked with two objectives: first, to introduce any information that has not been possible to annotate automatically (direct speech?, page breaks?, tables?), and second to correct possible mistakes during automatic annotation. At this moment, exactly what information will be annotated has not been defined yet. See Doc2.

Open question. Manual revision and annotation are also complex and time-consuming tasks. We must decide how they will be done.

6 Step 5. Publication.

XML files must be publicly available in a specific repository. WG4 is working to define the appropriate repository for the corpus.

Once a valid XML will be obtained, it will be published at the official repository. Therefore, the corpus will be available during review/annotation process.

The repository should guarantee that the novels could be downloaded as XML files and as plain texts. We can also demonstrate the use of more sophisticated technologies such as TEI-Publisher (teipublisher.com/) or txm (textometrie.org).

We can offer a DOI for each novel and for the whole corpus through Zenodo: *https://zenodo.org/*.

7 Step 6. Evaluation.

Two aspects must be evaluated: text quality and annotation.

Text quality could be evaluated by checking if text samples follows the criteria of Doc1. The idea here is to find recurrent mistakes. The annotation of the structure of each novel could be evaluated in the same way, but in relation to Doc2. Metadata of each novel must be manually checked in order to ensure the correctness of data.

Each file must be a valid XML-TEI document according to the defined XML scheme. XML can be validated with standard tools as oXygen, Jing, XMLlint, the TEI-Validator (http://teibyexample.org/xquery/TBEvalidator.xq) or any other tool.

Ideally, manual annotation must be evaluated following a double annotation by two annotators and then calculating inter-annotators agreement. This standard approach will show us the quality of the annotation scheme and the annotation process. It is not clear at this moment which information will be manually annotated. Therefore, we suggest to leave this point for later (linguistic annotation).

8 Bibliography

[1] Leech, Geoffrey (2005) "Adding Linguistic Annotation" in Wynne, M (editor), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Available online from http://ota.ox.ac.uk/documents/creating/dlc/

[2] Stubbs, Amber and James Pustejovsky (2012) *Natural Language Annotation for Machine Learning*, O'Reilly.

A APPENDIX 1. List of catalogues.

- OCLC WorldCat *https://www.oclc.org*
- Gutenberg catalog (in RDF) http://www.gutenberg.org/wiki/Gutenberg:Feeds

B APPENDIX 2. List of digital repositories

- Gutenberg project [several languages]: http://www.gutenberg.org/
- Oxford Text Archive (OTA) [several languages]: http://ota.ox.ac.uk/
- Deutsches Textarchiv (DTA - German Text Archive) [German]: http://www.bbaw.de/en/research/dta
- TextGrid Repository [German]: https://textgrid.de/en/digitale-bibliothek
- OBVIL Bibliothèque [French]: http://obvil.paris-sorbonne.fr/bibliotheque
- ARTFL-FRANTEXT [French]: http://artfl-project.uchicago.edu/content/artfl-frantext
- Biblioteca Virtual Miguel de Cervantes [Spanish (Castilian, Catalan and Galician)]: http://www.cervantesvirtual.com/
- Liber Libri https://www.liberliber.it/online/ [Italian]
- Biblioteca Digital Camões http://www.cvc.instituto-camoes.pt/conhecer/biblioteca-digital-camoes/literatura-1.html [Portuguese]
- Digitale Bibliotheek voor de Nederlandse letteren http://www.dbnl.org/ [Dutch]
- National Library of Poland http://www.bn.org.pl/en/digital-resources/polona/ [Polish]

C Appendix 3. Tools and resources for digitizing texts.

- https://www.digitisation.eu/
- http://www.impact-project.eu/index.php
- https://www.digitisation.eu/tools-resources/tools-for-text-digitisation/
- http://www.impact-project.eu/taa/tech/tools/