A Cor infrastructure for textual analysis – From Woolf to Verne

Nicholas Hayward
ancientlives@gmail.com
Loyola University Chicago, United States of America

Introduction

The Woolf Online project, which recently completed its second phase of development at Loyola University Chicago, sought to address the following fundamental questions about the nature and development of literature:

- how does a literary text come into being?
- what kinds of influence are at work upon the writer during the process ofinitial composition, and thereafter?

The Woolf Online project sought to investigate various ways in which different recoverable histories of a particular text could be used to illuminate the process of its composition. By recording the history of a particular text, its cultural, political, and autobiographical contexts and their interaction, visually we began to answer some of the following important questions:

- how is textual history related to other histories of a text?
- what use does literary criticism make of textual and contextualhistories?

Publication of Digital Scholarly Editions

As part of the development of the Woolf Online project, we developed an extensible development and publication framework, called 'Mojulem', for editing, publication, and visualisation of digital scholarly editions. We are now continuing this development with the Verne Digital Corpus, which focuses upon the work of Jules Verne, including original French language editions and their myriad, often questionable, English language translations.

Mojulem allows us to build on the concept of 'knowledge sites', as suggested by Peter Shillingsburg (2006), supplementing a core publication framework with modules/plugins such as OCR, editors, and image viewers.

Mojulem also enables us to host multiple projects within one installed framework, thereby enabling cross-project research, where applicable, and the option to aggregate specified data. Development of Mojulem, with the Woolf Online project and Verne Digital Corpus as examples of the current ongoing working environments, initially followed the need for four underlying core structures. These structures include CorPix, CorTex, CorCode, and CorForm, which are detailed as follows.

CorPix

Manuscripts and printed texts materially unite the iconic and lexical, the autographic and allographic, whereas all digital representations separate these constituent elements into images and transcriptions. With Mojulem projects, the common default display reunites the image and transcription by mapping the one to the other at the pixel level. CorPix software currently includes eHinman Transparent, TransparentOCR, Magnify, and Zoom. For example, pixel-level positioning and coordinate fixing is an inherent feature of both Transparent and TransparentOCR, within both editor and visualisation tools. eHinman is a digital adaptation of the original Hinman collator, and enables fade from the image of a page from one copy to another, thereby enabling a visual collation of multiple copies. Transparent is used within both the visualisation and editing stages of a project's development, enabling an editor and user alike to view the image as the primary entry consideration for the project.

CorTex

The CorTex is the stable resource containing the merged or compacted plain text transcriptions of the variant expressions of a work. It stores all information about text and variations, ready to be extracted for display of variation amongst versions; it is not necessary to recompute them.

The CorTex is the entity to which all standoff properties (markup, annotations, links, etc.) points, and on whose stability the system depends. It is as the source of each version's text and variation from other texts. The stability and endurance of the CorTex is protected by multiplying duplicate copies locked with a digital signature, which verifies for each user that a CorTex copy is viable. Analysis of the CorTex variable forms provides statistical feedback to guide the production of a conflated text, for example with the English language translations of a given Verne edition. Whilst

these statistical results are no guarantee of an ultimately correct translation, they offer a conflated text with the highest viable agreement amongst the provided collated texts. Textual disagreements are currently resolved by assigning probability values, a higher value defining a greater probability of accuracy and agreement amongst the collated texts. Using such probability results, we are currently able to filter problematic passages in each translation to conflate a text with the highest probability of agreement amongst the translations per edition.

These results can then be provided for further research and assessment, and act as a suitable starting guide for further analysis of the conflated text, and translation in the example of Verne's text.

CorCode

CorCode is the add-on value of analysis, argument, and explanation. Mojulem stores markup separately, as standoff properties, applying it as the user invokes it for the rendering of a specific item's image or text within a given visualisation, such as a transparent view of a page of the Initial Holograph Draft of 'To the Lighthouse'. To do this, Mojulem includes an editor which saves text and encoding separately, and filters for converting legacy, code-embedded transcriptions, including TEI encoded documents, into separate forms with markup analysed into properties, and filters for reversing this process.

CorForm

A CorForm[8] is a CSS stylesheet, containing special formatting rules, used to transform the overlapping properties of the CorCode into HTML. Each CorCode has a default CorForm, but other CorForms can be used in combination or as alternatives. Since a CorTex may have many CorCodes, and each CorCode many CorForms, structuring or formatting of the text can be attained by specifying some combination of already available resources, or by supplying new ones.

The CorPix, CorTex, CorCode, and CorForm are aggregated for a project within the Mojulem framework. Each such item is identified by a unique key, which is used as an index into the repository or database.

In addition to the initial four cores, identified above, we have begun development of CorAssess, allowing effective assessment and analysis of Cor data for the Verne translations and conflated English language texts.

CorAssess

CorAssess works in tandem with the CorTex to provide analysis of statistical and end results relative to the conflated text output. CorAssess allows us to visualise where text has been conflated based upon resolved disagreements, the points of disagreement and resultant probabilities between collated texts per edition, variance between collated texts, and visualise alternative resolution patterns relative to variation distance for given points of disagreements in the conflated text. With Verne texts, for example, we will be able to visualise conflation decisions, and offer alternatives for given decisions and disagreements, where applicable, in our conflated English language translations.

Why Verne?

After Woolf Online, we chose to focus upon the corpus of Jules Verne, including original French language editions and English language translations. We have begun collecting, collating, and preparing digitised copies of as many digitised editions as extant online. We have also been digitising early editions to provide an ever-growing dataset of Verne material.

The nature of early English language translations of Verne's editions is a continuing source of frustration for those interested in the works of Jules Verne. His early categorisation as a predominantly children's author in English language countries, unlike the publishing by Hetzel, coupled with early restricted access to original French language editions, simply compounded the issue.

The corpus of Jules Verne offers an interesting opportunity for literary and contextual analysis coupled with data processing and automated analysis. The myriad existing digitised English language translations, including US and British variant editions, often more prevalent than their counterpart, original French editions in our current digitised corpus, allows us to examine the development of those texts by comparing agreements, disagreements, omissions, and continuing revisions in said translations since a novel's first edition. We are hoping to use this analysis to filter the noise of years of collective translations to collate a unified English translation for each French language edition.

We will then be offering a comparison of French language edition against a filtered, collated English language edition. This will allow further consideration of the requisite merits of the English language translation directly juxtaposed to the original French language edition.

Conclusion

The development and combination of the initial four cores, CorPix, CorTex, CorCode, and CorForm, within the modular and adaptable framework Mojulem, allowed the second phase of the Woolf Online project to begin to approach the fundamental questions about the nature and development of literature, as briefly outlined in the introduction. With the addition of CorAssess, we are now beginning to address additional issues with the publication, transmission, and development of texts. We are also testing, and proving, the viability of Mojulem beyond the Woolf Online project.

The corpus of Jules Verne provides a particularly fascinating opportunity to test these cores, and provide a resultant conflated, English language translation per extant French language edition.

This paper will briefly introduce the Mojulem framework and its initial four cores, grounded in the example of the Woolf Online project, and detail the ongoing developments to augment this work with the above new work on the corpus of Jules Verne, and the ongoing Verne Digital Corpus

Bibliography

Shillingsburg, P. L.(2006) "From Gutenberg to Google: Electronic Representations of Literary Texts." Cambridge University Press.

Goodman, N. (1976) "Languages of Art." Hackett Publishing.

Hayward, N. & Shillingsburg. P.L. (2013) *Woolf Online*. Loyola University Chicago.. http://www.woolfonline.com