PROGETTO MAMBRINO
HISTORIAS FINGIDAS

Early printed edition and OCR techniques: what is the state-of-art? Strategies to be developed from the working-progress Mambrino project work

Tiziana Mancinelli
(Universität zu Köln)

Abstract

Some thoughts of a work-in-progress digital edition project. Limits and advantages of OCR (Optical Character Recognition) techniques. The use of Transkribus.
Keywords: Transcription, digital scholarly edition, OCR, Digital Humanities, Mambrino Project

In questo breve testo sono riportate riflessioni e considerazioni riguardo alcune fasi del percorso di realizzazione dell'edizione digitale del progetto Mambrino. Verranno discussi i limiti e i vantaggi dell'utilizzo di software per l'OCR (Optical Character Recognition). In particolare, cercheremo di riportare il lavoro compiuto su Transkribus.
Parole chiave: Transcrizione, facsimile, edizione digitale, OCR, Progetto Mambrino, Informatica umanistica

§

This short paper attempts to sketch out some findings on the digitization process of texts included in the Mambrino project (http://www.mambrino.it). The project is named after Mambrino Roseo da Fabriano who wrote translations of a huge corpus of Italian chivalry romances from Spanish literature, which were published in Venice during the Renaissance, between 1544 and 1565. These long novels were quite widely read at the time –only the Amadis de Gaule cycle includes more than twenty novels of about 800 pages each– and, as such, made a significant contribution to the knowledge of the European Renaissance, not only in the literary but also in the historical and socio-cultural realms. The Mambrino project will publish digital scholarly editions of these early printed books, through a visualisation of Cinquecentina facsimiles and transcribed text, accompanied by indexes for recovering minimal structural metadata.

Curating a digital scholarly edition is a long process: each phase requires human intervention that is both time-consuming and expensive. As such, we are forced to work slowly. Even the first phase, the mere digitisation of printed material, already requires a series of acknowledgements. High-resolution images play a very important role in the preservation of historical documents, as they are sometimes

Tiziana Mancinelli, «Early printed edition and OCR techniques: what is the state-of-art? Strategies to be developed from the working-progress Mambrino project work », *Historias Fingidas*, 4 (2016), pp. 255-260. DOI: 10.13136/2284-2667/65. ISSN 2284-2667.

«better than the originals». They can also help to widen access to the data and will save time and work as the project progresses.
As Patrick Sahle writes «digitised edition is not a digital edition» (2016, 26): capturing an image is the first step towards building a digital scholarly edition but it is not sufficient: «That is why digitisation may change the accessibility of a printed edition and may add at least some basic functionalities such as searching - but digitisation does not make a printed edition a digital edition» (27).

Molecules in configuration[1]

Some of the books have already been digitised and are now available on our website in high resolution facsimiles. Images can be converted into text using optical character recognition (OCR), software or keyboarding and later encoded using standard mark-up languages for web searching and retrieval of text. Although printed books pose fewer problems than those that are hand-written, mid-sixteenth century printed books have proved quite difficult to transcribe because the characters used cannot be processed very easily with a simple OCR. The 16th Century editions are highly irregular and, although the writing is in italic and almost always well printed, often manual printing (attached letters, abbreviations, errors, misalignments) or the material history of the book (stains, use, rips, transparencies) makes the OCR (Optical character recognition) process very complicated and with many, leading to multiple errors. Manual transcriptions are uneconomical and impractical, whereas a high resolution image could allow a text to be amenable for a OCR.

However, as many corpora of all kinds of books during the Renaissance, in several languages, in the whole European sixteenth and seventeenth century, used the so-called «Aldino» font –italic type– based on a standard form of calligraphic handwriting –the OCR recognises the text with a high percentage of errors. Aldino font was designed by Francesco Griffo in Venice obeying to the directions of Aldus Manutius. These amazing characters were imitated by all the most important printers in Europe. Therefore, training a software on this particular font would be a very useful project for a huge corpora of books.

Transkribus and Fine Reader: the playground of OCR. Academic project and proprietary software? Limits and advantages[2]

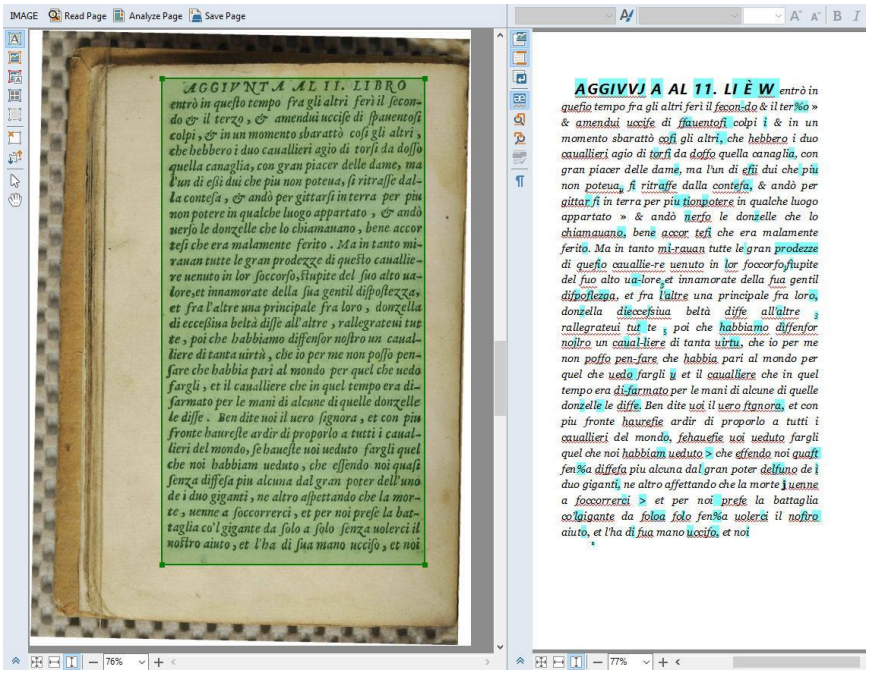
In this paragraph we are going to present our choices and strategies for the OCR work. Firstly we tried «Fine Reader» a common proprietary software that

[1] «Written texts, at the basic level of documents, consist of molecules (usually paper or ink) in configurations conforming to a semiotic or sign system (often an alphabet) arranged according to some rules of deployment (grammar)» (Shillinsburg 2006, 57).
[2] The last paragraph came from Alberto Bazzacco's contributions and screenshots. Some parts were directly translated from his notes. Without his work and his effort, we could not achieve this evaluation.

provides accurate text recognition. And lately we discovered «Transkribus» a transcription and recognition platform for handwritten texts[3].

The development of those kind of OCR software for different kinds of documents – in particular handwritten documents - has given us hope that we can speed up the transcription process. Our collaborator, Alberto Bazzaco, has made a very valuable effort to achieve good results using OCR. Below an analyses and an evaluation of his outcomes.

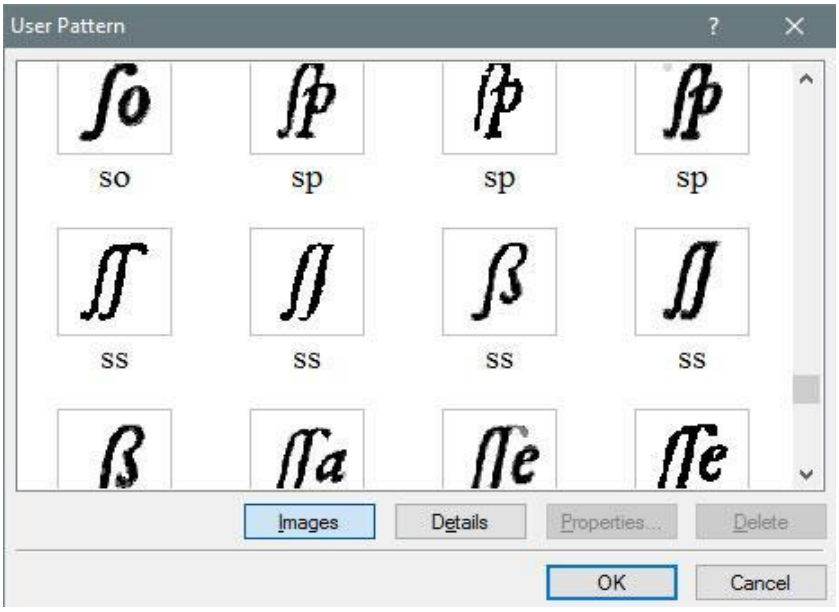
The first step was to feed the machine with the original version. Fine Reader resulted in a very high number of errors. In particular, some characters, such as tironiana notes and ties, were completely wrong:



Using this software at this stage, would mean that the manual labour to create a transcript would be enormous.

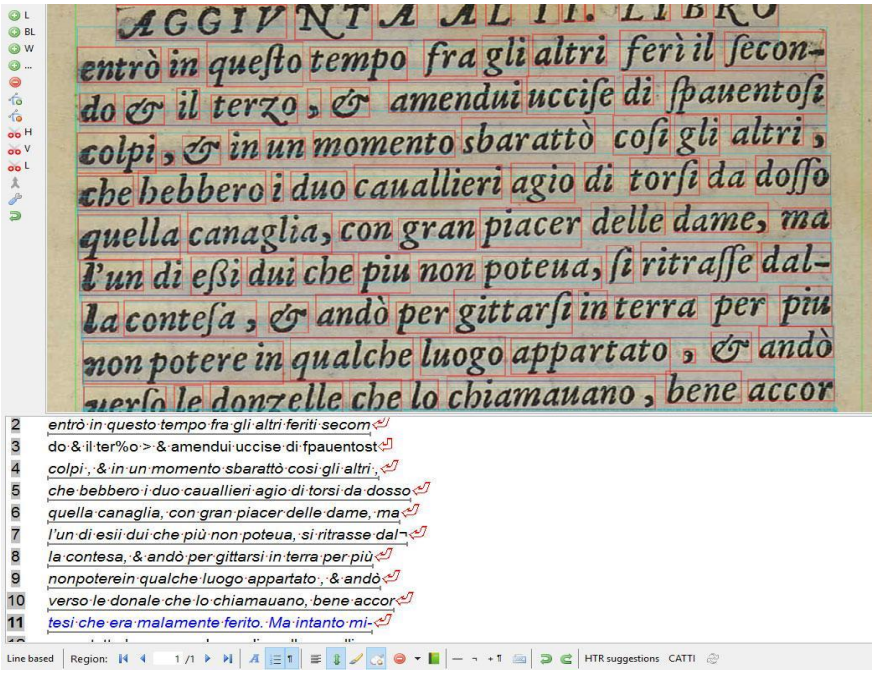
So the software was then trained: it consisted of giving Fine Reader some identified characters. After running it for about twenty pages, it created a «pattern»:

[3] Transkribus URL < https://transkribus.eu/Transkribus/ > (Accessed: December 20, 2016)



Having two other full transcriptions of this corpus of books, in order to achieve a good result, Alberto also decided to introduce a custom dictionary and to automatize the corrections. The results were much more acceptable.

The Transkribus tool, a transcription and recognition platform currently used for handwritten texts also needs to be trained to recognize handwritten texts of a certain author only with a repeating operations and make the machine learning a writing. Transkribus (TS) was good enough from the beginning. Transkribus provides helpful tools to define the text regions, its lines and their baselines.



This software provides a frame for transcription and forces us to understand the reading order of the letter. When reading a letter, we are often faced with interline additions. But whereas humans intuitively integrate those additions into the reading flow, a programme like Transkribus lacks such an intuition. What a transcriber can do to keep the reading flow linear is to integrate the interline additions by inserting extra lines and baselines for them (Jander, 2016).

There are some advantages and disadvantages of working with these different tools. Fine Reader can work locally, and you can upload high-resolution pictures without needing a good connection. It also allows you to export in many formats (including epub, html, odt) and use a custom user dictionary. The «find & replace» is also applicable to all pages but not in one click. In Transkribus find» is possible using regular expressions, but not «replace» –replacements have to be made by hand. Above all, it doesn't allow you to launch a series of regular expressions with just one click. And you cannot upload a personal dictionary. The program has already trained a routine for the recognition of ancient prints, and the result is immediately noticeable. Images and text appear in the same window so that comparison is facilitated. The program engine has no difficulty in recognizing lines and words. You can export to TEI (Text Encoding Initiative)[4] and the exported file also indicates the spatiality of words in the image (this option is not of interest to our project), but I point out why it might be useful for the creation of PDF-MRC. Still, also Transkribus, does not allow to launch a series of «find & replace» with just one click.

The improvement and development of a OCR software does not sort out yet the high percentage of errors that one could find while using it. However, there are new opportunities as a promising project as Transkribus could give us new goals for our project. Although it is still at a early stage of development. Our work on Mambrino project suggests that starting a large-scale project developing OCR *ad-hoc* for the Aldino font would be a very interesting opportunity for many different scholars.

§

[4] TEI (Text Encoding Initiative) www.tei-c.org

References

Bognolo, Anna; Cara, Giovanni; Neri, Stefano, *Repertorio delle continuazioni italiane ai romanzi cavallereschi spagnoli. Ciclo Amadis de Gaula,* Roma, Bulzoni, 2013

Jander, Melina, «Handwritten Text Recognition – Transkribus: A User Report», *The electronic Text Reuse Acquisition Project (eTRAP)*, Institute of Computer Science, University of Göttingen, Germany, 2 November 2016. URL: < http://www.etrap.eu/transkribus-a-user-report/ > (Accessed: December 20, 2016)

Pierazzo, Elena, *Digital scholarly editing: Theories, models and methods.,* Aldershot, Ashgate, 2015.

Rydberg-Cox, Jeffrey A, «Digitizing Latin Incunabula: Challenges, Methods, and Possibilities», *Digital Humanities Quarterly*, 3/1 (2009). URL: < http://www.digitalhumanities.org/dhq/vol/3/1/000027/000027.html > (accessed: December 20, 2016)

Sahle, Patrick, «what is a Scholarly Digital Edition?», in *Digital Scholarly Editing Theories and Practices*, ed. Matthew James Driscoll and Elena Pierazzo, Cambridge, Open Book Publishers, 2016.

Shillinsburg, Peter, *From Gutenberg to Google: Electronic Representations of Literary Texts: From Gutenberg to Google: Electronic Representations of Literary Texts*, Cambridge, Cambridge University Press, 2006.