Digital edition from a mixed corpus: early printings and manuscripts of Martianus Capella

This work deals with getting several early samples digitized, both samples printed and handwritten, of the same ancient Latin text, De nuptiis Philologae et Mercurii by Martianus Capella, an author from the Fifth Century. The aim is to build a digital critical edition of the text with the possibility of comparing all the variations of the same base text, which should be as close as possible to the text used during the Renaissance.

The main task is to digitize several early samples, both printed and handwritten, of the same ancient Latin text: De nuptiis Philologae et Mercurii by Martianus Capella.

Martianus' text is an encyclopedic work from the Late Antiquity concerning the Seven Liberal Arts with frequent quotations of both Latin and Greek sources, divided into nine books. The corpus considered for that work, as shown on the upper row, is the set of the first four complete printed editions.

But Vulcanius’ Earlier 1577 process informs in the preface, that B. Amerbachius lent him some manuscripts to prepare his edition.

It is very likely that the two manuscripts of Martianus Capella preserved in Basel University Library were used by Vulcanius and possibly also by H. Petrus, former Martianus’ editor in Basel (1532). Therefore, both manuscripts, FV17 from the 16th century and FV40 from 1490 have also been included in the corpus.

The modern critical editions have taken as a base text the Grotius edition text, which became the vulgata for centuries.

Grotius then used his master previous edition, and so on and so forth until the princeps.

Early prints and manuscripts of the same text have obvious differences, but both share some features. Early prints were presented to the public as a combination of the common manuscripts in the early decades of the print. Both share resembling shapes, materials, page layouts, abbreviations, ligatures, etc.

The OCRing of the corpus

Long Short Term Memory (LSTM) Neural Networks have been checked in historical prints with very good results. The most tested OCR engines are ABBYY, Tesseract, and OCRopus.

I have chosen the open-source systems OCRopus for OCRing. I have also applied the tools and the methods and proposed by Springman et al. belonging to CIS in LMDE. OCRopus was built on a sample set for training with complete line segmentation. The OCRing of the corpus. There major issues: 1) the page layout recognition, 2) the separation of texts and paratexts, and 3) the transcription of historical spellings. Since OCRopus system works segmenting the text in lines, some additional issues are: Page layout recognition. Separation of texts and paratexts. Transcription of historical spellings.

Next step was to train the system and then predict the text. The editions took up to 36,000 iterations for training the system with the results that you can see on the table. In the case of manuscripts I yielded as good output for FV17 as for editions, but much worse for FV40, as expected. For FV17 the line segmentation was made with relative few mistakes. 20,000 training iterations were made and the OCR results were similar or even better than the ones made for the editions. Despite some manual cleaning of some pages in FV40, the line segmentation provoked many mistakes. However, I typed 13 pages with the marginal and interlinear glosses when included in line image. In the training process many of the badly segmented lines remained blank. Then, I made a much longer training: more than 200,000 iterations, but such a hard-starting point yielded poor results.

Conclusions, Prospects and Enhancements.

The results of OCRing depend on many factors:

- Quality of images is crucial. Not just the resolution, but the state of original material (stains, etc).
- Preprocessing images is also important to obtain good line segmentation.
- Choosing the set for training with complete representation of every kind of script and character.
- Postprocessing and profiling with PoCoTo and Profiler with the complete set of historical and characters.

Manuel Ayuso