ENCODE Workshop 'AI and ancient writing cultures', Bologna 23-27 gennaio 2023

Report by Benedetta Bocchi (University of Parma)

The activities of the Encode Workshop "AI and ancient writing cultures" opened on January 23rd with a methodological and an ontological introduction led by **Gioele Barabucci** (Norwegian University of Science and Technology) to the nature of Artificial Intelligence and the methods based on AI that find application in the field of Digital Humanities.

Gioele Barabucci, animated by a shared aversion for the unaware use of tools, which "allow you to obtain results but cancel the thought", started from a (relatively) simple example that involves the conversion of an image representing the roman numeral IV in machine-readable and reproducible information. He then introduced the processes that occur within the so-called "black box": to execute this operation, various methods can be applied, including the "vertical edge detection", which requires the use, in the specific case, of the Sobel operator. Actually, there are multiple operators to perform the same operation, as for example the Prewitt operator. The use of a given operator instead of another one is determined by the loss function, which calculates the accuracy of one over the other. This practical example was followed by the introduction to a specific vocabulary, including the definitions of methods, tools, libraries, dictionaries and frameworks.

The second part of the session concerned the methods of representation of knowledge and of information, according to the two main schools of thought "symbolic" and "subsymbolic". The representation is followed by a series of operations:

- Embedding, which is the same as spatial orientation.
- Classification, which we can perform by writing an algorithm.
- Clustering, which requires the subdivision of groups of elements arranged in the multidimensional space through embedding.
- Forecasting, which consists of the automatic generation of new data starting from those already supplied to the machine, now particularly à la page after ChatGPT went "viral".

The second and the third sessions of the morning were run respectively by an epigraphist, **Aaron Hershkowitz** (Institute for Advanced Study, Princeton), and by a papyrologist, **Nicola Reggiani** (University of Parma). **Aaron Hershkowitz**, in his presentation entitled "Signal and Noise: Epigraphic Ventures in Machine Learning", presented several projects that involve the application of machine learning to epigraphic documents:

- ITHACA: a project focused on Greek epigraphic texts. The project deals with three macro-problems: the restoration of fragmentary texts, their geographical location and their chronology.

- Classifying Latin inscriptions of the Roman Empire: a project that focuses on Latin epigraphic texts. The aim of the project is to transfer the epigraphic typologies of the Epigraphik Datenbank Clauss-Slaby (EDCS) into the categories of the Epigraphic Database Heidelberg (EDH) which refer to the controlled vocabularies of EAGLE.

- Automated writer identification: a project based on an attempt to automate S. Tracy's ability to match inscriptions with the same stonecutter for dating purposes.

- Reconsidering the Roman Workshop: a project that involves the usage of images of Latin inscriptions and which, therefore, similarly to the previous one, exploits computer vision technologies. The main purpose of the project is to reconstruct and date the inscriptions through the analysis of the layout of the writing on the epigraphic support.

- KRATEROS: a project on digitization of epigraphic squeezes in 2D and 3D at the Institute for Advanced Study in Princeton, coordinated by Aaron Hershkowitz himself.

Thereafter, **Nicola Reggiani**, offered an overview of the use of Artificial Intelligence applied to the study of papyri with a lesson entitled "The Artificial Papyrologist at Work - Digital Papyrology and the AI". He started provocatively with the reading of the reply generated by ChatGPT to the following request: "Tell me about the Artificial Intelligence in papyrology". He continues then with an excursus on the evolution of the professional figure of the papyrologist, from the "digital" to the "artificial" papyrologist, followed by the mention of some projects that involve the application of AI to the papyrus support.

The afternoon of this first day of the Workshop was entirely dedicated to the training run by **Audric Wannaz** (University of Basel) "Analyze Tabular Data: some Tools and Workflows". Assuming that combining Artificial Intelligence and Humanities often means being able to manipulate tabular data, he presented three different tools that allow to analyse some datasets, which he himself provided (Greek and Latin papyrus fragments, the first two books of the *Aeneid* and the Euripidean *Iphigenia*). The following tools were presented:

- Jupyter Notebook, which allows you to import csv files and write codes in Python.
- Orange Data Mining, a toolkit that enables data analysis, machine learning and data mining.
- Streamlit, a Python framework for building web apps.

The core lesson that trainees have learnt was that different tools may be more or less suitable for our type of research. What matters, even in choosing the right tool for us, is its compatibility with the research question.

The second day (January 24th) of the workshop was almost entirely dedicated to training activities. The first training, entitled "Annotating papyrus images for the paleographer and for AI: What, how and why?", was led by **Isabelle Marthot-Santaniello** (University of Basel). First of all, she introduced the D-Scribes project, whose purposes are reuniting fragments, identifying scribes and characterising scripts. The D-Scribes project is based on the concept of machine learning since, in order to pursue the already mentioned objectives, it is necessary to “teach” the machine, that is to train the machine. For example, to go in the direction of automatic recognition of the scribes, it is necessary to train the machine by uploading materials of certain attribution (due to the presence of signatures or on a palaeographic basis). Only after having "fed" the machine with enough material, in order to evaluate the effectiveness of the model that we have thus come to establish, we need to question the authorship of papyri that haven’t been uploaded yet, but whose scribes we obviously know. And here the work of the papyrologist is truly irreplaceable: a computer speaks numerically and returns numbers and percentages, but the interpretation of these statistical results is up to the scholar.

To go back to the training, the task was to make the transcription of a papyrus (a Coptic papyrus fragment of the second book of *Deuteronomy*) through the READ platform and to identify the individual letters and the various forms of the individual letters, which, for example, can be useful for attributing a text to a certain hand/scribe.

The second training, entitled "Automatic semantic analysis of Ancient Greek", was held by **Alek Keersmaekers** (KU Leuven). During this training participants were guided in the use of R, a programming language oriented towards statistics, and R Studio, an editor for R. In particular, the exercises, carried out under the guidance of the trainer, focused on tabular data imported into R Studio and their subsequent manipulation necessary to answer any possible research question. For example, a research question could concern the recurrence within a corpus of papyri of the name of a specific profession. From the numerical response of the machine, the scholar can draw conclusions about the appearance or, on the contrary, the disappearance of a profession at a given chronological height or identify the chronological range of maximum diffusion of a profession.

The afternoon of the second day was dedicated to Transkribus: the group workshop organised by **Pietro Liuzzo** (Bibliotheca Hertziana - Max Planck Institut) was, in fact, followed by the presentation of a project that uses the platform and directly involves the University of Bologna, MemoBo. During the group work of **Pietro Liuzzo**'s session, participants were given a scanner and a 4-page document in English and printed script. By downloading Transkribus (the expert tool version) or using Transkribus Lite (the online version), trainees scanned the pages of the printed document with a special app and uploaded them in jpeg format. On the image, the various areas that present the writing were selected, the text transcribed in the lower part of the screen and divided into numbered lines. The transcript was saved and a model suitable as possible for our type of text was selected. After training the machine with the transcription of a single page, participants tested the results that the machine gave for the other 3 pages, which were automatically analysed by the program. A discussion between the trainer and the trainees followed, with an evaluation on the type and frequency of errors, numerically consistent, as expected given the small transcription inserted. A final debate involving the four different working groups closed the session, aimed at bringing out the pros and cons of this system and the peculiar characteristics of the recurring errors in the transcription performed automatically by the machine.

The last session of the day was run by **Edward Loss** (University of Bologna) and was entitled "Challenges and issues of using Transkribus in large late mediaeval manuscript collections: The Memoriali Project (MemoBo))". The speaker introduced us to the world of the mediaeval *Memoriali* and to the MemoBo project. The *Memoriali* are contracts between private individuals stipulated in the city of Bologna and its surrounding countryside with a value greater than 20 lire of *bolognini.* These documents were registered by notaries in the so-called *Libri Memorialium* and covered a period of about two hundred years. In these documents there are also poetic attestations in dialect drawn up by notaries, including one of the earliest evidence of the circulation of Dante's texts. After this introduction, **Edward Loss** illustrated the pros but also the critical points encountered by his working group in using Transkribus and in developing a model for the automatic transcription of these manuscripts.

The training process of the machine, indeed, requires the selection of manuscripts that stood out for their clarity and rigour, both of the scripts themselves and of the support. The first memorial selected with these criteria was the one of Enrichetto delle Querce due to the clarity and correctness of the writing. The speaker stressed that it is necessary to teach the machine also the abbreviated

forms contained therein, in order to make a good diplomatic transcription. In addition to the transcription, the text is also tagged to distinguish proper and common nouns, including those of various professions. After three attempts made on the aforementioned memorial, the MemoBo team arrived at the elaboration of a good model which automatically transcribes Enrichetto's manuscript but, when they tried to apply the same model to the autograph specimens of another scribe of the same period and with the same preparation, the latter seemed not to work as well. We concluded that it is necessary to develop different models for different scribes. This is the *status quaestionis* at the moment as the research of the MemoBo working group is still in progress.

The third day (January 25th) started with a lecture by **Margherita Fantoli** (KU Leuven), entitled "Automatic tagging and parsing of Latin texts: methods, tools and challenges". After a brief mention of the Index Thomisticus Project, one of the first products of the Digital Humanities, **Margherita Fantoli** illustrated the steps necessary to process a proposition and, more in general, a text. To pursue this aim, the following steps are essential:

- Tokenization, which provides the subdivision of character strings into minimal units (tokens).
- Lemmatisation, which consists in returning the words to their basic form (lemma).
- Part-Of-Speech (POS) Tagging, which allows to tag the function of the single word within the sentence.
- Morphological Features, which attributes morphological characteristics to single tokens.
- Syntactic tree, which reconstructs the syntactic structure of a proposition in the form of a syntactic tree.

These steps are mandatory for any project involving the annotation of texts. What changes within the specific projects is the style of the annotation, which varies according to the conventions. It is therefore necessary to be aware of the existence of several conventions, which differ in crucial respects, and to follow one and only one of these conventions throughout the entire annotation work.

After these indications of general and methodological nature, participants were introduced to a list of some projects for the annotation of Latin texts, some of which are based on manually annotated corpora. The reasons for annotating texts can be several, such as the study of the characteristics of the language in a diachronic and synchronic sense, the analysis of the peculiar features of an author's language but also educational purposes. However, annotating a large number of texts manually can be hard and time-consuming, and thus came the possible use of solutions for processing natural languages, which are different, but all characterised by methods based on Artificial Intelligence. The three alternatives to manual annotation are:

- Symbolic/Rule-based NLP.
- Statistical/Machine Learning NLP (which can be applied in a supervised system or not; the supervised one is the most widespread and the one which Fantoli focused on, explaining the machine training methods and the evaluation of the results returned through the confusion matrix method).
- NLP with Deep neural networks.

The lesson ended with a practical session: participants tried to use an existing text annotation tool, named UDPipe. By uploading one of the Latin texts made available by **Margherita Fantoli** (extracts from Caesar's *De Bello Gallico* and Petronius' *Satyricon*) trainees let the tool generate the

morphological analysis of the text and the syntactic tree and subsequently calculated the accuracy degree of the results and evaluate the type of errors.

The remaining part of the morning was dedicated to the training titled "Basics of Python and Jupyter Notebooks: Q&A". The training involved carrying out, independently or in pairs, the exercises at the bottom of the lessons of the basic Python course, followed online in the weeks before the workshop. During the activities, the support of **Margherita Fantoli** has been fundamental, also for the introduction to the several Python libraries.

The afternoon session was run by **Andrè Walsøe** (Medida) and **Andrea Gasparini** (University of Oslo) and was entitled "A practical introduction to machine learning and natural language processing on papyrus data". After a brief introduction aimed at illustrating the various possibilities that open up by applying Machine Learning and NLP technologies to written production, such as the classification of a text into predefined categories, its translation, the tagging of its elements and the automatic generation of a text, the training focused on the joint application of NLP and ML technologies to papyrus data. The activity was carried out using Google Colab with pre-prepared Python scripts useful for navigating the previously provided papyraceous data. The lines of Python code already prepared by Walsøe have been started but also modified in order to obtain different information from the uploaded data.

The fourth day (January 26th) was entirely dedicated to the conference "Artificial Intelligence and Ancient Writing Cultures".

1. In the first paper entitled "Artificial Intelligence and the Palaeography of Greek and Coptic papyri: potential and limits", **Isabelle Marthot** presented the D-Scribes project to a wider audience. D-Scribes aims to reunite papyrus fragments, to identify scribes and to characterise writings. This project had previously been presented to the trainees during the second day of the workshop but, during the conference, **Isabelle Marthot**'s speech was enriched by **Stephen White** (University of Venezia), one of the developers of the READ (Research Environment for Ancient Documents) platform.
2. The presentation of **Silvia Ferrara** (University of Bologna), entitled "State-of-the-art AI technology applied to the decipherment of ancient scripts: Results and prospects" set out goals, purposes and achievements of the INSCRIBE (Invention of Scripts and their Beginnings) project. The project reconsiders the inventions of writing and is dedicated, in particular, to the study of three still undeciphered scripts of the Aegean area (Cretan Hieroglyphic, Linear A and Cyprus-Minoan), going beyond the traditional methods of cataloguing inscriptions. INSCRIBE aims to publish a digital corpus of the inscriptions of the three aforementioned scriptures together with 3D models of artefacts, accompanied by a multidimensional interface that tags the inscriptions, the types of inscribed objects, the provenance, the function and the archaeological context.
3. **Mark Depauw** (KU Leuven) gave a lecture entitled "AI and extracting information from primary sources and secondary literature", offering an overview of the new horizons of Trismegistos which since 2008 has been collecting metadata on texts from all over the ancient world applying systems based on Artificial Intelligence and, later, also on machine learning. In particular, since last year Trismegistos has also been working on the so-called secondary literature developing a tool that allows to extract the references of ancient documents and authors from the texts and convert them into the Trismegistos standard, providing the abbreviations with an ever-increasing number of links.

4. After the coffee break, **Hussein Mohammed Adnan** (University of Hamburg) presented a paper entitled "Beyond Textual Content: Analysing Patterns in Written Artefacts", introducing the public to the analysis, through computer vision technologies, of written artefacts from the perspective of a computer engineer and anticipating what would have been the core of the following day of the workshop.
5. Thereafter, **Charlotte Tupman** (University of Exeter), in her presentation "Next steps in analysing the layouts of inscribed texts" illustrated the Reconsidering the Roman Workshop project already mentioned by trainers of the workshop the previous days. The aim of the project is to analyse the methods of drafting and layout of the texts on the epigraphic support. The project aims to use machine learning methodologies to examine the processes that lie behind the drafting of the registered texts. To this aim, **Charlotte Tupman** and her team at the University of Exeter use a technology based on neural networks applied to image processing. Thus, once the characters have been located in an image, their regularity in size, shape, spacing, position and orientation, and the overall shape of their outline can be analysed. Investigating the layout of the inscriptions is not only interesting from a historical point of view for the reconstruction of the working methods of the ancient stonecutters but can also be useful in reconstructing the fragmentary inscriptions precisely through the use of technologies that process results in a "predictive" way.
6. At the end of the morning was **Gioele Barabucci**'s presentation "AI in research: unfounded fears and serious risk", rhetorically elaborated in a *pars construens* and a *pars destruens* and spiced up with an irony that reveals the unfounded fears and concrete risks hidden behind the application of technologies AI-based to research in general and, more precisely, to research in the humanities. On the one hand, there are widespread unfounded fears and prejudices that hinder the spread of AI, such as the one whereby Artificial Intelligence will end up replacing papyrologists, paleographers, philologists and other scholars who deal with ancient writing cultures. On the other hand, however, there are real risks in the use of AI, which can be summarised in two points: the high costs required by projects that involve the use of AI and the danger that research will fail due to an abandonment to usage of tools that are easier to use. The latter danger, though, can be faced through the combined application of several methods which require the collaboration of computer scientists and humanists.

The final morning (January 27th) was entirely dedicated to a training (or rather a dialogical lesson) held by **Hussein Mohammed Adnan** and entitled "The Pattern Analysis Software Tools (PAST)". Starting from the assumption that the manual analysis of manuscripts, papyri and more generally ancient artefacts that contain writing generally requires a lot of time and can be subject to human errors, **Hussein Mohammed Adnan** underlined the need for scholars to benefit from the rapid progress occurred in different fields of Artificial Intelligence in order to facilitate them in a more efficient study of written artefacts and help them answer their research questions. Then, he moved on to the introduction to PAST: the Pattern Analysis Software Tools (PAST) is a set of software tools, developed at the Center for the Study of Manuscript Cultures (CSMC) for the automatic analysis of visual and tabular patterns research data originating from the study of ancient written artefacts. The application of these tools to them can open up new research horizons, stimulated by the statistical approach that these tools apply to research data already processed by scholars. The tools included in PAST and elaborated by **Hussein Mohammed Adnan**, are the following:

- Handwriting Analysis Tool (HAT): HAT is a software tool that can be used to analyse multiple and different handwriting styles. A similarity score can be calculated for each predefined style to create a relative comparison between them with respect to an unknown style.

- Visual-Pattern Detector (VPD): VPD is a software tool for pattern detection. This tool can be used to automatically recognize and allocate visual patterns (such as words, drawings and seals) in digitised manuscripts.

- Line Detection Tool (LDT): the main goal of the LDT is to analyse images of writing supports in order to detect lines (such as papyri fibres) and estimate their density. These detected lines form a pattern, which can be used as a distinctive feature of the writing support.

- Text-Lines Counter (TLC): this software tool is able to detect, count and mark the text lines in images of handwritten manuscripts. The latest version of TLC also contains features to detect vertical text lines and bright text lines on a darker background.

- X-ray Fluorescence Data Analysis Tool (XRF-DAT): the main goal of the XRF-DAT is to analyse tabular data generated by X-ray Fluorescence (XRF) spectroscopy in order to ease and speed up the processing and evaluation of data obtained when analysing written artefacts; in particular their inks, pigments and writing supports.

- Artefact-Features Analysis Tool (AFAT): the main goal of this tool is to calculate statistical information from manually generated tabular data which consists of distinctive features of artefacts. The value of each feature in these tables is represented by a positive integer, which describes the particular variant of this feature in an artefact. Furthermore, each variant can have any number of variations, represented by an alphabetical letter. The combination of variant and variation can be used to describe the observed version of a given feature in an artefact.

After this exhaustive explanation, space was left for questions, starting a debate on the nature and role of professional figures such as the one of the "digital humanist". **Hussein Mohammed Adnan** seemed to have a clear position in this regard: digital humanities are not only the product of the digital humanist but the result of the collaboration between computer scientists and humanists. The two categories of experts will certainly have to come together and approach each other's disciplines and vice versa, but the elaboration of research questions remains the exclusive task of the humanist, to whom the computer scientist can provide tools for finding answers. Here it's also necessary to recall the position of **Gioele Barabucci**, for whom the professional figure of the digital humanist is fundamental for his "double nature": the digital humanist should be the *trait d'union* between the antiquity sciences and the digital world.