

'Channels of Digital Scholarship' Seminar

New tools and old questions in the analysis of textual corpora

Research in the digital humanities has experienced explosive growth and development in the last ten years. Two important factors have contributed to this progress: firstly, the very strong mobilisation of scientific and scholarly communities to engage with this emerging field in all humanities sectors; secondly, the extraordinary progress of digital tools and capacities. This has resulted in a profusion of initiatives at all levels: major digitisation projects led by libraries and academic institutions, digitisation of corpora of all kinds and for all periods, and multiple research projects with targeted objectives.

The aim of this first Channels of Digital Scholarship seminar series is to reflect upon new avenues for the analysis and use of textual corpora. Textual corpora and their uses represent several challenges in the development and validation of digital tools for analysis, the dialogue between disciplines, and the institutional structures that support the wide range of projects that are being developed. In this series of four seminars, the Maison Française d'Oxford and Digital Scholarship @ Oxford, with the help of leaders of digital humanities initiatives in the CIVIS network, propose to explore these challenges from Franco-British and international perspectives.

Convenors: Goran Gaber (EHESS (LIER-FYT)), Andrew Cusworth (Digital Scholarship, Oxford), Christophe Gaillac (Nuffield College, Oxford), Pascal Marty, Olivier Delouis, Tristan Alonge (MFO), Grégoire Lacaze (MFO / Aix-Marseille Université, LERMA).

SEMINAR PROGRAMME (UK time)

24 May, 2-4pm: 'Institutions & Networks: The cultural infrastructures of digital scholarship'

25 May, 2-4pm: 'Greek and Latin corpora'

31 May, 2-4pm: 'From the Renaissance to the Enlightenment'

1 June, 2-4pm: 'Building and Mining Corpora for Social Media Discourse Analysis'

Institutions & Networks: The cultural infrastructures of digital scholarship

14:00-16:30, 24.5.2022
Seminar Room, Maison Française d'Oxford & Zoom

Convenors: Goran Gaber, Andrew Cusworth

In this unique two-hour seminar hosted by Maison Française d'Oxford (MFO) and Digital Scholarship @ Oxford (DiSc), participants from leading research institutions in France and the UK will discuss the challenges and opportunities for national institutions engaged in digital research, including

- creating and sustaining long-term infrastructures in a research environment that favours short-term funding;
- enabling collaboration and exchange;
- and balancing public and commercial interests in their work.

Introduced by the directors of MFO and DiSc (Professors Pascal Marty and Howard Hotson), this seminar will draw together speakers from Bibliothèque Nationale de France, Huma-Num, and the Alan Turing Institute. Gathering in the seminar room at MFO, participants and attendees will have an opportunity to hear four papers on these topics, take part in formal and informal discussion and exchange, and enjoy a little cheese, wine (or alcohol-free alternative), and conversation.

Speakers:

Nicolas Sauret (Huma-Num CNRS): “The HN Lab of Huma-Num. Prospects and initiatives around AI for the humanities”

Emmanuelle Bermès & Marie Carlin (BNF, Paris): “Digital humanities at the BnF: between age-old missions and support for new uses”

Barbara McGillivray (King’s College, London): “Mining for meaning in the open: computational analyses of historical textual corpora”

Dave De Roure (Engineering Science Department & Alan Turing Institute): “Knowledge Infrastructure for Digital Scholarship”

IN-PERSON REGISTRATION LINK:

<https://oxford.onlinesurveys.ac.uk/mfo-registration>

ZOOM REGISTRATION LINK:

<https://us06web.zoom.us/join/ztEudOCrpjlvHtI5fHad7LS7zg74LrHQY2hr>

Greek and Latin corpora

14:00-16:00, 25.5.2022
Zoom

Convenors: Olivier Delouis, Christophe Gaillac, Tristan Alonge

The study of Greek and Latin languages is ever more concerned with corpora analysis. Collections of texts have developed dramatically over the last twenty years. Almost all Greek ancient and medieval literature from Homer to the Fall of Constantinople in 1453 is today digitized through the “Thesaurus Linguae Graecae” or TLG (by the University of California, founded 1972), while on the Latin side the “Brepolis’ Library of Latin Texts” offers an immense array of texts from the beginnings of Latin literature until the present day (by Brepols Publisher, founded 1991). Inventories of these classic corpora, including growing collections in open access, are regularly made and enable studies that are generally limited to each individual case (see for instance *Digital Classical Philology, Ancient Greek and Latin in the Digital Revolution*, dir. Monica Berti, 2019).

Now, there are many methods and tools applicable to the analysis of modern languages. Still, the branch of artificial intelligence that helps computers to understand human languages, i.e. Natural Language Processing (NLP), remains underdeveloped for classical languages. Many concepts used in modern corpora analysis such as deep learning-based approaches, convolutional and recurrent neural networks, contextual language models or recently bidirectional encoder representations from transformers (BERT) are still far away from being used in classical humanities.

In this seminar, we aim to present the work of scholars who engage in cross-disciplinary collaboration between the study of classical literature and NLP.

Speakers:

Thibault Clérice (École nationale des chartes, PSL, Paris) – “Detecting sexual isotopies in Latin corpora: setting up an experiment and first results”

Marianne Reboul (IHRIM – UMR 5318 & ENS Lyon): “Homer and Machine Learning: translations alignment on Iliad and Odyssey”

Thea Sommerschild (Università Ca’ Foscari Venezia) – “Working with Greek epigraphic data for Machine Learning”

REGISTRATION LINK: https://us06web.zoom.us/meeting/register/tZYudO-vrjsqH9f5d_XFlyfqGHa-NLdH_8Sg

From the Renaissance to the Enlightenment

14:00-16:00, 31.5.2022
Zoom

Convenors: Goran Gaber, Tristan Alonge

For several reasons, Early Modernity represents a fascinating field for Digital Scholarship. On the one hand, the abundance and variety of written and printed material – ranging from academic treatises, reference works, and newspapers to maps, theatre registers, and private correspondence – offer almost infinite possibilities for historical inquiry. On the other hand, the finite nature of these textual corpora presents scholars with a reasonably delimited and thus practically manageable area of research. Last but not least, widespread intellectual interest in this period regularly results in sustained large-scale projects of digitisation, interdisciplinary and institutional collaboration, as well as technical and scholarly innovation.

It should therefore be of little surprise that such a conjecture has given rise to a large number of ground-breaking projects that have not only presented “old material in a new light” by, for example, processing, encoding, and analysing historical texts but have irrevocably altered the landscape of Early Modern scholarship as such, by enlarging both our understanding of what counts as historical material, as well as the scope of questions that such material can answer. The third session of the seminar series will thus present cutting-edge research initiatives from both sides of the Channel dealing with different types of textual corpora from the Renaissance to the Enlightenment.

Speakers:

Nicholas Cole (The Quill Project, Oxford): “The records of negotiation: problems and opportunities”

Nicholas Cronk-Glenn Roe (The Voltaire Foundation, Oxford): Title TBC

Howard Hotson (Cultures of Knowledge, Oxford): “Did Hartlib have a Circle? New Methods for Answering Old Questions”

Maria Susana Seguin (IHRIM – UMR 5318 & Université Paul-Valéry Montpellier - IUF): “Constituting a virtual Corpus: the case of Philosophie [cl@ndestine](#)”

REGISTRATION LINK: <https://us06web.zoom.us/meeting/register/tZEvd-qhrDMjHtA8w5ZqY5GUnWUICa1e9v19>

Building and Mining Corpora for Social Media Discourse Analysis

14:00-16:00, 1.6.2022
Zoom

Convenor: Grégoire Lacaze

Social media discourse analysis raises the topical question of the process of building a corpus of digital posts. The determination of the limits of the corpora is at stake in this process. In this round table, we will discuss the amount and types of data that need to be selected in the building of a corpus.

Digital platforms of social media have the main property to be regarded as open environments in which new posts and comments can be added without a limitation in time, which has a strong impact on the singularity of corpora that can be elaborated at a given time.

The question of reproducibility applied to this data according to the FAIR principles (Findable, Accessible, Interoperable, Reusable) will also be tackled. Once the corpora are constituted, they have to be stored on safe and permanent repositories, which directly leads us to highlight the importance of open data for long-term analyses.

When the corpora are built, they can be analysed thanks to data-mining techniques. Different approaches and methodologies will be presented, some of them being based on deep learning techniques including neural networks. Digital corpora obviously need digital tools to be analysed. Algorithms and software such as open source Iramuteq will be shown.

A recurrent question as far as corpus building is concerned is the dichotomy between qualitative analysis and quantitative analysis.

Speakers:

Bernie Hogan (Oxford Internet Institute, Oxford): “Theorising and integrating platform signals into digital text corpora”

Gudrun Ledegen (Université Rennes 2, Laboratoire PREFICS): “Suicide prevention chat, quantitative and qualitative description of a discourse genre for better listening”

REGISTRATION LINK:

<https://us06web.zoom.us/meeting/register/tZ0oduuqrj8jGtaykEVmZnxXQHT96nCArhR0>