CHR2023 Workshop:
AI and Large Language Models (LLMs) for the Analysis of Large Literary Corpora

December 5, 2023

Venue

The workshop will take place at the Ecole normale supérieure, in salle Dussane, 45 rue d'Ulm, 75005 Paris, France. It is held in coordination with the CHR 2023 Conference (Dec 6-8, 2023, EPITA, Paris).

Registration is **mandatory** at this link: CHR2023-Workshop registration.

The workshop will be on site. Remote attendance will be possible: a link will be send the day before the workshop to participants who registered with the link above.

Situation

The availability of large collections of literary texts (several thousands of novels for a given language for example, covering a significant part of the literature of the time) along with statistical models have profoundly changed our knowledge of literature. In parallel, the availability of efficient natural language processing (NLP) tools has made possible the structural analysis of these novels.

More recently, the advent of large language models and more specifically generative AI has again dramatically modified the analysis of literary texts, providing more robust and more versatile annotation tools. Zero-shot learning means that new categories and new tasks can be explored at a reduced cost, through prompting for example. But this is not without raising new questions. These techniques may be less robust (depending on the quality of the training set), harder to evaluate and harder to replicate (since models evolve very quickly; they depend on several parameters and do not always produce the same output).

The workshop will explore themes relates to the annotation and analysis of large literary corpora . It will more specifically examine for what generic tasks we now have access to relatively robust and accurate tools. We will then investigate to what extend generative models can be exploited in this context, their benefits and their potential drawbacks. The implication on teaching may also be addressed, as well as the very quick obsolescence of current programs, given the pace of the evolution of the domain.

Schedule

- **9:45-10:00: Introduction.**

- **10:00-10-45: The Promise and Peril of Large Language Models for Cultural Analytics**
  David Bamman (Berkeley, USA).

- **10:45-12:00: Analyzing Large French Literary Corpora with Fr-BookNLP**
  Frédérique Mélanie, Jean Barré, Olga Seminck, Thierry Poibeau (CNRS & ENS/PSL, France):

- **Lunch**.

- **1:30-2:15: Prediction and Surprise**
  Ted Underwood (Illinois Urbana-Champaign, USA).

- **2:15-3:00: Automatic Information Extraction from Literary Works for Audiobooks Generation**
  Elena Epure (Deezer, France) & Gaspard Michel (Deezer & Loria, France).

- **Break**.

- **3:30-4:15: Computationally Modeling Collective Narratives**
  Andrew Piper (McGill, Canada).

- **4:15-5:15 Debate: LLMs, Generative Models and Literary Analysis: where are we going?**

Sponsors

With the support of Lattice (https://lattice.cnrs.fr), CNRS (IRN Cyclades) and Prairie (Paris Artificial Intelligence Research Institute, https://prairie-institute.fr).

Scientific committee

David Bamman (Berkeley, USA), Evelyn Gius (Darmstadt, Germany), Thierry Poibeau (CNRS, France), Sara Tonelli (FKB, Italy).

Organization committee

Jean Barré (jean.barre@ens.psl.eu), Pedro Cabrera, Florian Cafiero (florian.cafiero@sciencespo.fr), Fabien Garrido, Virginie Pauchont, Marie Puren (marie.puren1406@gmail.com), Thierry Poibeau (thierry.poibeau@ens.psl.eu).