CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

TNA ACTIVITY REPORT

Programmable corpora as linked data

Author: Federico Pianzola

Current position: Assistant Professor

Affiliation: University of Groningen

Host institution: University of Potsdam

Mentor(s): Peer Trilcke, Ingo Börner

Period of stay: 27th March 2023 – 21st April 2023

Introduction

Working on the ERC Starting Grant project "Graphs and Ontologies for Literary Evolution Models" (GOLEM, https:// golemlab.eu) (Pianzola et al., 2023), I developed a graph database of fiction corpora taken from various online sources in 5 different languages. The goal is to describe texts using derived data referring to various textual features, so that comparisons between texts can be done without accessing the full text of the stories. This is an optimal solution to make linguistic and literary information of in-copyright material available to many users. The idea is similar to the HathiTrust Extracted Features dataset (Jett et al., 2020), but the features encoded as metadata are much richer and also refer to narrative and stylistic elements relevant for reader response (e.g. characters, relationships, topics, readability, etc.).

The infrastructure and tools of the DraCor project (Fischer et al., 2019), together with the concept of "programmable corpora", are an excellent framework to visualize the metadata of the stories in the GOLEM database and provide some ready-made analysis regarding characters' networks. In drama and DraCor, characters are a central element and that is also the case for most of the text in the GOLEM database: fanfiction stories created mainly on the basis of the affective attachment that authors have for characters. During the fellowship, I adapted some of the DraCor tools so that they could be used for the needs of the GOLEM project but also for other projects interested in giving access to programmable corpora with complex metadata stored in a knowledge graph.

Methodology

The goals of the fellowship have been pursued in three steps:

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

1. create a data model to describe narrative and fictional texts in a structured way according to schemas that are general enough to be reusable for other project and allow interoperability (e.g. CIDOC-CRM, LRMoo, CLS-INFRA ontology);
2. adapt the DraCor web frontend so that it displays information relevant for narrative and fictional texts;
3. designing an API that could feed the DraCor frontend starting from metadata stored in a knowledge graph according to the data model developed in step 1 (cf. Börner & Trilcke, 2023).

Results

Thanks to the collaboration with the DH Potsdam team, I have been able to complete all three steps and achieve the proposed goals.
The adapted web frontend is available in the following repository: https://github.com/GOLEM-lab/golem-corpora-frontend.
The API is available in the following repository: https://github.com/GOLEM-lab/golem-frontend-api. The GOLEM API has been conceived as a Python middleware between the web frontend, which needs JSON files as input, and data stored in a knowledge graph. The API itself uses endpoints similar to that of the DraCor API and can also be used with other programming languages to retrieve data from the knowledge graph (out of the box, it works with OpenLink Virtuoso Open Source edition 7). In addition, the Python middleware offers more advanced functionalities and a modular structure that allows to send customized SPARQL queries to the knowledge graph database.

The data model has been implemented only to the extent needed for the functionality of the main frontend page (Figure 1). The basic classes and properties included so far are about the segmentation of a text into documents and paragraphs, the wordcount, the number and gender of the characters featured in a story, and the number of readers' comments available for a story. An example of the RDF modelling of this information can be seen in the following file:
https://github.com/GOLEM-lab/golem-frontend-api/blob/main/data/manual_example_data.ttl.

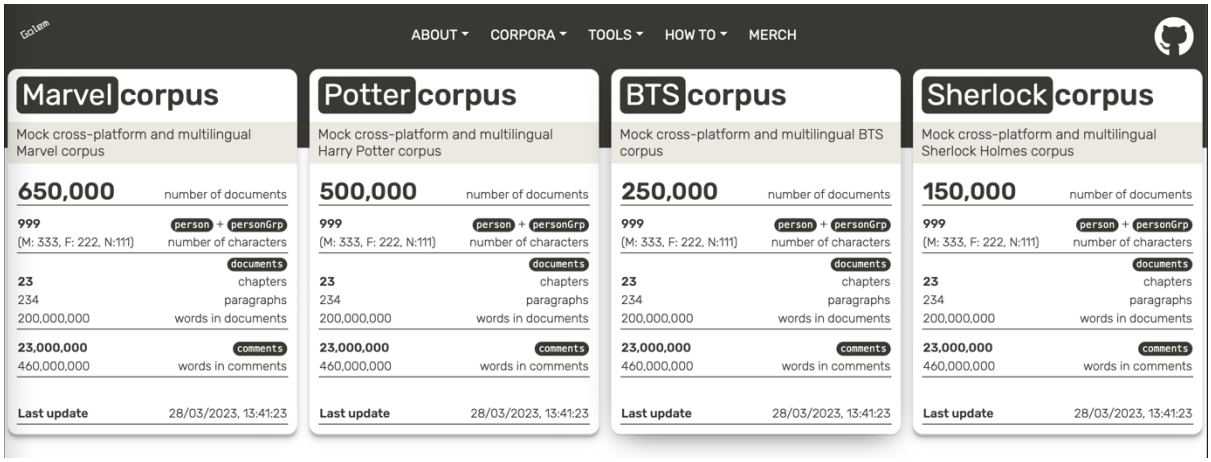

Figure 1. Screenshot of the adaptation of the DraCor web frontend.

Impact and future work

During the fellowship I have grasped the software architecture of both the API and the React web frontend, learning how to modify them to add new modules for the retrieval of more information from the knowledge graph and its display in the web frontend. The flexibility of the Python API also allows to create customized SPARQL queries and implement them as Python methods that can be more easily used by users with no knowledge of SPARQL. In this way,

we hope to give more users the chance to use the GOLEM data without the need to acquire additional skills, since the output of the Python methods are JSON files that can be easily used as such or converted into a dataframe with one line of code.

Adapting the DraCor frontend and API will extend the reusability of the DraCor resources beyond texts encoded in TEI, allowing to use them also with in-copyright material and other corpora for which only derived data are available. Additionally, the current and future development of the GOLEM ontology in compliance to the CIDOC-CRM ISO standard and the CLS-INFRA data modelling schema, will ensure the potential reusability and interoperability of the GOLEM infrastructure.

Future work in the GOLEM project will involve the further development of an ontology for narrative and fiction, with the addition of more fine-grained derived data that describe characters, events, stylistic features, narrative strategies, and reader response. Accordingly, the API and web frontend will be also extended to allow for the retrieval and display of such information.

The impact of this fellowship is extremely high for the Computational Literary Studies community, encouraging the sharing of derived data and facilitating the interoperability of analytical tools with various corpora, file formats, and storage solutions.

References

Börner, I., & Trilcke, P. (2023). *CLS INFRA D7.1 On Programmable Corpora*. https://doi.org/10.5281/ZENODO.7664964

HTRC Derived Datasets—Documentation—HTRC Docs. (n.d.). Retrieved 28 September 2021, from https://wiki.htrc.illinois.edu/display/COM/HTRC+Derived+Datasets

Pianzola, F. (2020). Linked-Potter: An example of ontology for the study of the evolution of literature and reading communities. In B. Hodošček (Ed.), *JADH2020 Proceedings of the 10th Conference of the Japanese Association of Digital Humanities "A New Decade in Digital Scholarship: Microcosms and Hubs"* (pp. 28–32). Graduate School of Language and Culture, Osaka University. https://jadh2020.lang.osaka-u.ac.jp/programme/longpaper/pianzola-linked.html

Pianzola, F., Acerbi, A., & Rebora, S. (2020). Cultural accumulation and improvement in online fan fiction. *CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands*, *2723*, 2–11. http://ceur-ws.org/Vol-2723/short8.pdf