Constructing the GOLEM: Graphs and Ontologies for Literary Evolution Models

Pianzola, Federico
f.pianzola@rug.nl
University of Groningen

Yang, Xiaoyan
xiaoyan.yang.21@alumni.ucl.ac.uk
University of Groningen

Visser, Noa
noavisser@proton.me
University of Groningen

van der Ree, Michiel
michiel.van.der.ree@rug.nl
University of Groningen

van Cranenburgh, Andreas
a.w.van.cranenburgh@rug.nl
University of Groningen

This paper presents the first release of a graph database of online fiction corpora taken from various sources in five different languages (English, Spanish, Italian, Indonesian, Korean). The goal is to describe texts using "derived data" (OECD 2005) – or "mesodata" (Boot 2009) – referring to various textual features, so that comparisons between documents could be done without accessing the full text of the documents. The idea is similar to that of the HathiTrust Extracted Features dataset (Jett et al. 2020), but the features encoded in the GOLEM project ("Graphs and Ontologies for Literary Evolution Models") are much richer and also refer to narrative and stylistic elements and to reader response data (e.g. characters, relationships, topics, readability, sentiment of comments received by the story, etc.) collected from likes and comments left on the stories. Something similar has already been done on a smaller scale for a selection of texts in English (Piper 2022) and Dutch (Luoto and van Cranenburgh 2021). The creation of the GOLEM has been inspired by such work but will operate on a completely different scale, which requires the automation of the extraction of textual features for millions of stories.

The core concept of the GOLEM infrastructure is that of "programmable corpora", i.e. "research-oriented corpora providing an API" (Fischer et al. 2019), which allows to easily reapply scripts, notebooks, and pipelines of analysis to all texts in the corpora, inasmuch as they are encoded following the same principles and can be queried via the same API and SPARQL endpoint. Since the GOLEM focuses primarily on derived data, there is no need for a resource-intensive XML database of texts encoded in TEI. Only statements about the texts and their reception will be stored in the database, following existing ontologies as closely as possible in order to maximize the compatibility with other relevant projects, like Wikidata and MiMoText (Schöch et al. 2022).

Beside choosing three of the most spoken Western languages, Korean and Indonesian have been included because these cultures have a peculiar role in the worldwide digital reading landscape. On one hand, K-pop and K-drama inspire many works of fanfiction in all the mentioned languages, and are quite influential among youth, particularly in Europe. On the other hand, the Indonesian case is extremely important to understand the evolution of fiction more broadly, because it is culturally very distant from all the other considered countries (Muthukrishna et al. 2020) and Korean culture is very influential among Indonesian youth. Hence, it will be interesting to compare how cultural traits spread differently in countries with different wealth, educational level, and cultural influence. Moreover, Indonesia is a densely populated developing country in which for many people it is easier to access online fiction than print books (Rokib 2019; Yoesoef 2020) thus it offers a lot of data.

Once metadata and derived data for all texts will be included, the GOLEM will be an almost complete database of all fanfiction published online in English, Spanish, Italian, Indonesian, and Korean during the years 2000-2022, which correspond to almost the whole lifetime of the genre of online fanfiction. As such, the GOLEM will be an excellent resource to test hypothesis about the evolution of fiction writing and reader response without the influence of external factors like the historical selection by publishers or educational curricula. We have already created a pilot ontology (Pianzola 2020) and used it for some analyses (Pianzola et al. 2020), showing how this kind of data can offer interesting insights for research in both cultural evolution and literary studies.

https://golemlab.eu

Bibliography

Boot, Peter (2009): *Mesotext: Digitised Emblems, Modelled Annotations and Humanities Scholarship*. Amsterdam: Amsterdam University Press.

Fischer, Frank / **Börner, Ingo** / **Göbel, Mathias** / **Hechtl, Angelika** / **Kittel, Christopher** / **Milling, Carsten** / **Trilcke, Peer** (2019): "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama", in *Digital Humanities 2019: "Complexities" (DH2019)*, Utrecht. DOI: 10.5281/zenodo.4284002.

Jett, Jacob / **Capitanu, Boris** / **Kudeki, Deren** / **Cole, Timothy** / **Hu, Yuerong** / **Organisciak, Peter** / **Underwood, Ted** / **Dickson Koehl, Eleanor** / **Dubnicek, Ryan** / **Downie, J. Stephen** (2020): "The HathiTrust Research Center Extracted Features Dataset (2.0)". DOI: 10.13012/R2TE-C227.

Luoto, Severi / **van Cranenburgh, Andreas** (2021): "Psycholinguistic Dataset on Language Use in 1145 Novels Published in English and Dutch", in: *Data in Brief* 34. DOI: 10.1016/j.dib.2020.106655.

Muthukrishna, Michael / **Bell, Adrian V.** / **Henrich, Joseph** / **Curtin, Cameron M.** / **Gedranovich, Alexander** / **McInerney, Jason** / **Thue, Braden** (2020): "Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance", in: *Psychological Science* 31 (6): 678–701. DOI: 10.1177/0956797620916782.

OECD (2005): "Derived Data Element", in: *OECD Glossary of Statistical Terms* https://stats.oecd.org/glossary/detail.asp?ID=5130 [04.11.2022].

Pianzola, Federico (2020): "Linked-Potter: An Example of Ontology for the Study of the Evolution of Literature and Reading Communities", in: Hodošček, Bor (ed.): *JADH2020 Proceedings*

of the 10th Conference of the Japanese Association of Digital Humanities "A New Decade in Digital Scholarship: Microcosms and Hubs". Osaka, Japan: Graduate School of Language and Culture, Osaka University: 28–32 https://jadh2020.lang.osaka-u.ac.jp/programme/longpaper/pianzola-linked.html [04.11.2022].

Pianzola, Federico / **Acerbi, Alberto** / **Rebora, Simone** (2020): "Cultural Accumulation and Improvement in Online Fan Fiction", in: *CHR 2020: Workshop on Computational Humanities Research*, November 18–20, 2020, Amsterdam, The Netherlands. CEUR Workshop Proceedings: 2–11 http://ceur-ws.org/Vol-2723/short8.pdf [04.11.2022].

Piper, Andrew (2022): "The CONLIT Dataset of Contemporary Literature", in: *Journal of Open Humanities Data* 8(0). DOI: 10.5334/johd.88.

Rokib, Muhammad (2019): "The Polemics of Digital Literature in Indonesia", in: *Advances in Social Science, Education and Humanities Research* 380: 287–292. DOI: 10.2991/soshec-19.2019.63.

Schöch, Christof / **Hinzmann, Maria** / **Röttgermann, Julia** (2022): "Smart Modelling for Literary History", in: *International Journal of Humanities and Arts Computing* 16 (1):78–93. DOI: 10.3366/ijhac.2022.0278.

Yoesoef, M. (2020): "Cyber Literature: Wattpad and Webnovel as Generation Z Reading in the Digital World", in: *Proceedings of the International University Symposium on Humanities and Arts (INUSHARTS 2019)*. Depok, Indonesia: Atlantis Press. DOI: 10.2991/assehr.k.200729.025.