

Event annotation for literary corpora analysis

Claude Grunspan, Frédérique Mélanie-Becquet, Jean Barré, Laurette Chardon, Ioana Galleron, Marco Naguib, Clément Plancq, Olga Seminck, Thierry Poibeau

▶ **To cite this version:**

Claude Grunspan, Frédérique Mélanie-Becquet, Jean Barré, Laurette Chardon, Ioana Galleron, et al.. Event annotation for literary corpora analysis. Digital Humanities 2022, ADHO, Jul 2022, Tokyo, Japan. pp.471-472. hal-03738806v2

HAL Id: hal-03738806
https://hal.science/hal-03738806v2

Submitted on 15 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Event annotation for literary corpora analysis

Publication presented at the DH2022 conference (Digital Humanities Conf.), ADHO, Tokyo (held online)

Claude Grunspan[1] *(claude.grunspan@sorbonne-nouvelle.fr)*
Frédérique Mélanie[1] *(frederique.melanie@ens.psl.eu)*
Jean Barré[1] *(jean.barre@chartes.psl.eu)*
Laurette Chardon[2] *(Laurette.chardon@unicaen.fr),)*
Ioana Galleron[1] *(ioana.galleron@sorbonne-nouvelle.fr)*
Marco Naguib[1] *(marco.naguib@hotmail.com)*
Clément Plancq[3] *(clement.plancq@univ-tours.fr)*
Olga Seminck[1] *(olga.seminck@cri-paris.org)*
Thierry Poibeau[1] *(thierry.poibeau@ens.fr)*

[1] *Lattice (ENS-PSL & CNRS & Université Sorbonne nouvelle)*

[2] *Crisco (Université de Caen*

[3] *MSH Val de Loire (Université de Tours*

Studying large corpora in the literature domain, especially novels, mean that new tools are needed in order to address narratological questions at scale. A large body of research has specific developed techniques for the task, giving birth to the field known as distant reading (as opposed to close reading, by a human being), (Moretti, 2013). In this paper, we present a series of tools providing the basis for the large-scale and comprehensive annotation of French novels through the adaptation of the BookNLP project (Bamman et al. 2014) to French. We present the

different kinds of annotation provided and then address specific issues concerning the annotation of events (Vauth *et al.*, 2021).

1. Event annotation within the BookNLP project

The BookNLP framework (Bamman et al. 2014) is one of these software ensembles integrating various modules (entity recognition [1], coreference [2], event and quotation analysis [3]) that can be applied to large collections of text. The initial BookNLP contained tools for English only, and a new project is now extending the range of languages covered. We are on our side developing the same kind of modules for French.

Natural language processing is now almost exclusively based on machine learning techniques, which means most of the effort required to develop this kind of tools lies in text annotation. For French, we have annotated 20 extracts of French novels from the 19 th and 20 th century. We build on the Democrat project [4], whose aim was to annotate a large corpus of French texts (from different historical periods and different genres) with coreference information. We selected the texts corresponding to our criteria (copyright free texts from novels from the 19 and early 20 th century), hence our 20 extracts (for a total of 184.000 words).

The task first consisted in annotating entities following the BookNLP guidelines and mapping the initial Democrat coreference annotations to BookNLP. We then focused on event annotation, as this is one of the key features for distinguishing between author styles, but also for identifying specific episodes in a story, such as the fortune changes of the main characters, or the climax of a story arc.

However, we discovered that annotating events is slightly more difficult than annotating entities. In BookNLP (Sims et al., 2019, Bamman et al., 2019 and 2020), the definition of the notion of event is as follows: “The event layer identifies events with asserted realis (depicted as actually taking place, with specific participants at a specific time) -- as opposed to events with other epistemic modalities (hypotheticals, future events, extradiegetic summaries by the narrator)”. The definition entails that verbs with a negation or with a modal are not annotated, for example, and only conjugated form of the verbs are annotated.

2. The necessity to integrate modals and negation in the annotation scheme

We chose to annotate all kinds of events, without the initial limitations imposed in BookNLP. The first example presents three sentences with approximately the same meaning. If we leave apart the conjugated verb in the main clause, all the sentences include another clause, with a conjugated verb in the first sentence (1a), with an infinitive in the second (1b), and with a participle in the third one (1c).

1a. *Après qu'il a mangé, il s'en est allé.*
1b. *Après avoir mangé, il s'en est allé.*
1c. *Ayant mangé, il s'en est allé.*
1d. *After he had eaten, he left.*

1a – 1c have roughly the same meaning and should thus be annotated with two events, independently of the form of the verb in the subordinate clause.

Negation is more complex, as *generally* a negation means that no event has occurred. But this is not always the case and examples like 2a can be found:

2a. *Il ne put retenir ses larmes.*
2b*. He could not hold back his tears.*

which roughly means that the character cried. In an example like this one, there is definitely an action so in our opinion it should be annotated as such. Here our choices differ slightly from the ones in the original BookNLP project.

All annotations were carried out after multiple rounds of discussions and the creation of a set of annotation guidelines heavily dependent on the initial BookNLP annotation scheme for events (but including the differences highlighted in this section). The total dataset comprises 14,305 events among 184,000 tokens in the 20 books in our corpus.

The annotated corpus as well as our guidelines are freely available on GitHub. A collection of computer programs makes it possible to go from our annotation to something close to the original BookNLP scheme by excluding from the corpus examples with a negation or a modal. The next steps will consist in evaluating the robustness of the developed solution and its ability to provide useful information for actual literary studies.

Bibliography

1. **Bamman D., Underwood T. and Smith N.** (2014). A Bayesian Mixed Effects Model of Literary Character, *Proceedings of the conference of the Association for Computational Linguistics* (ACL), Baltimore, USA, June 2014.
2. **Bamman D., Popat S. and Shen S.** (2019). An Annotated Dataset of Literary Entities. *Proceedings of the conference of the North American Association for Computational Linguistics* (NAACL), Minneapolis, USA, June 2019.
3. **Bamman D., Lewke O. and Mansoor A.** (2020). An Annotated Dataset of Coreference in English Literature. *Proceedings of the Language and Resource Evaluation Conference* (LREC), Marseille, France, May 2020.
4. **Landragin F.** (2021), Le corpus DEMOCRAT et son exploitation. *Langages* nº 224 (4/2021), pp. 11-24,
5. **Moretti F.** (2013), *Distant Reading*, Verso Books, London.
6. **Sims M., Park J.H. and Bamman D.** (2019). Literary Event Detection. Proceedings of the Conference of the Association for Computational Linguistics. Florence, Italy, July 2019.
7. **Vauth M., Hatzel H.O., Gius E. and Biemann C.** (2021). Automated Event Annotation in Literary Texts. CHR 2021: Computational Humanities Research Conference, Amsterdam, The Netherlands, November 2021.

Acknowledgements

This work was partially supported by the EUR ‘(Ecole Universitaire de Recherche) Translitteræ´ (programme “Investissements d’avenir” ANR-10- IDEX-0001-02 PSL* and ANR-17-EURE-0025). Thierry Poibeau is supported in part by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR19-P3IA0001 (PRAIRIE 3IA Institute). Lastly, this research also benefited from the support of the CNRS International Research Network Cyclades (Corpora and Computational Linguistics for Digital Humanities).

Notes

1. Person names, location names, etc.

2. A coreference occurs when two or more expressions refer to the same person or thing, like in ***Joe Biden*** $_i$ *said…* ***He*** $_i$ *was…* ***The president*** $_i$ *appeared to be …*

3. Roughly, who (the source) said what (the quotation).

4. https://www.ortolang.fr/market/corpora/democrat/3