A Corpus of Critical Citations Contexts

Frédérique Bordignon, Philippe Gambette

Journal of Open Humanities Data

A Corpus of Critical Citations Contexts

DATA PAPER

FRÉDÉRIQUE BORDIGNON
PHILIPPE GAMBETTE
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

We present here a corpus of 505 critical citation contexts, i.e. a set of sentences or propositions that contain at least one citation of a study towards which the author(s) has/have a negative opinion. Those contexts come from other existing annotated corpora, from our readings about critical citation and disagreement in science, and from contexts manually annotated by native speakers of English. We have re-annotated all those contexts in order to be sure that they match our definition of critical citations. This corpus can be helpful to train tools dedicated to the automatic retrieval of critical citations.

CORRESPONDING AUTHOR:
Frédérique Bordignon
Ecole des Ponts, Marne-la-Vallée, France; LISIS, INRAE, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France
frederique.bordignon@enpc.fr

KEYWORDS:
critical citations; negative citation; citation analysis; citation context; disagreement

TO CITE THIS ARTICLE:
Bordignon, F., & Gambette, P. (2024). A Corpus of Critical Citations Contexts. *Journal of Open Humanities Data*, 10: 39, pp. 1–6. DOI: https://doi.org/10.5334/johd.215

(1) OVERVIEW

REPOSITORY LOCATION

https://zenodo.org/doi/10.5281/zenodo.10694464

CONTEXT

This corpus of critical citation contexts was first built for the Cita&Re project (an exploratory project funded by Université Paris-Est (France) dedicated to the study of critical citations received by retracted papers) and then as part of the NanoBubbles project (an ERC Synergy funded project dedicated more broadly to the study of mechanisms of science correction). In both cases, this corpus allows us to conduct a linguistic study of critical citations in order to understand their mechanisms and to develop tools enabling their detection.

A citation is the explicit reference to another scientific piece of work within the full-text of a scientific publication. With a critical citation, authors carry a negative connotation towards the work they cite. In Bordignon (2022), we suggested critical citations can convey criticism (pointing out a weakness or a fault in the cited work), make a comparison (with the aim of expressing that one study is better than another) or question the cited work (conveying concerns, doubts, or uncertainty).

This is an extremely rare type of citation (Cano 1989; Catalini et al. 2015; Lin 2018; Oppenheim and Renn 1978; Spiegel-Rosing 1977) which complicates the constitution of a corpus dedicated to them, but which makes the corpus all the more valuable for the scientific community that is working on them.

(2) METHOD

This corpus derives from the results of three different methods aiming at gathering a set of contexts very likely to display critical citations. We then performed a new manual annotation to guarantee the quality of the corpus and ensure that the citations are critical citations according to the definition we have provided.

STEP 1 – RETRIEVING POTENTIAL CRITICAL CITATION CONTEXTS IN EXTANT CORPORA

With a literature review, we have identified existing corpora of citation contexts (i.e. sentences or snippets of text containing at least one in-text citation) that have been manually annotated. These corpora are generally shared by the computational linguistics community. We have identified six corpora providing for the polarity of the citation(s) (i.e. to what extent the author agrees with the paper they cite) occurring in each extracted context. As these corpora were built for different purposes and rely on different typologies of citation polarity, for each of them we have selected the contexts annotated with the label(s) that correspond to our definition of critical citation:

- The CONCIT corpus (Hernández-Alvarez, 2015), which consisted of citation contexts from the ACL Anthology Reference Corpus, i.e. conference and journal papers in natural language processing and computational linguistics. We have selected the contexts annotated with the “neg” label corresponding to the negative polarity (vs neutral or positive in her schema). The CONCIT corpus is CC BY-SA.
- The corpus shared by Athar (2011), along with his paper dedicated to the automatic identification of positive and negative sentiment polarity in citations to scientific papers. We have retrieved the contexts annotated with the “n” label corresponding to the negative polarity (vs objective or positive polarity in his schema). The corpus is distributed with no particular license but is made of ACL materials (CC BY-NC-SA or CC BY).
- The Citation Function Corpus has been made available by Teufel *et al* along with their paper (2006) aiming at automating the recognition of the rhetorical function of citations in scientific text, i.e. the author’s reason for citing a given paper. They proposed an annotation scheme relying on 12 categories/functions. We have retrieved contexts

annotated as "weak" (weakness of cited approach) and "CoCo-" (author's work is stated to be superior to cited work) in the XML "CFunc" attribute associated to each context. Those contexts are retrieved from conference articles in computational linguistics available from arXiv, an Open Access repository, and are distributed under the CC BY-NC.

- The DFKI corpus shared by Dong and Schafer (2011) with their paper presenting their automatic citation classifier. We retrieved contexts labelled as "negative" (vs "positive" and "neutral") in the "sentimental label" field. The authors present their dataset as public for research usage. All materials are drawn from the ACL Anthology corpus.
- The IMS corpus built by Jochim and Schütze (2012) to develop their citation classifier. They adopted the annotation scheme developed by Moravcsik and Murugesan (1975) combining different types of features. We therefore retrieved contexts bearing the "negational" feature (labelled as "CEPN", "CJPN" or "OJPN"). The corpus is distributed with no particular license but is made of ACL materials (CC BY-NC-SA or CC BY).
- *The corpus Ye et al (2020) built* for their study (Schneider et al., 2020) of the citations received by a particular retracted paper. We retrieved citation contexts annotated as "N" meaning "'Poor research' (negative)". The corpus is distributed under CC0 license.

STEP 2 – COLLECTING POTENTIAL CRITICAL CITATION CONTEXTS FROM RESEARCH ARTICLES

For the review work we did in a previous study (Bordignon, 2022), we have identified citation contexts given as examples by authors to illustrate what could be considered as critical citations, even if another wording is used (see the 56 different labels listed in Table 1 in this study). Our corpus (Bordignon, 2021) is distributed under CC BY license. Contexts are labelled as "Bordignon_2021" in the Source field and the "Source_Paper_ID" is the ID of the paper mentioning those examples (not the paper they are coming from, as this information is lacking from authors having identified and cited them as examples).

Table 1 Distribution of the critical citation contexts presented in the corpus.

SOURCE	# CONTEXTS
Athar_2011	221
Bordignon_2021	39
CFC	11
Concit	81
DFKI	24
IMS	43
Manually_collected_FB_PG	20
Pubmed_annotated_by_translators	63
Ye_et_al_2020	3
Total	505

Since then, we have continued to collect examples as we read. In the corpus, they are tagged "Manually_collected_FB_PG" and the "Source_Paper_ID" is the ID of the paper they occur in.

STEP 3 – RETRIEVING AND ANNOTATING CITATION CONTEXTS FROM PUBMED

As mentioned above, the constitution of the corpus we describe in this paper is part of the Cita&Re project, which aimed at investigating whether retracted or corrected articles received more critical citations than others. Therefore, we drew on the corpus of Hsiao and Schneider (2021) which contains contexts of citations received by 4,611 retracted articles, originating from 28,057 articles. In order to carry out the comparison that interested us at the time, we constructed a corpus comparable to the corpus of citation contexts of the retracted articles by recovering the citation contexts of the non-retracted articles also cited by these 28,057 articles. We retrieved the full-texts from PubMed Central, identified the citation contexts and then removed those mentioned in Hsiao and Schneider's corpus (i.e. those citing retracted articles).

DOI: 10.5334/johd.215

We built a corpus of 2,500 contexts by selecting one context out of 400. The Python scripts are available online (Gambette, 2024).

These 2,500 citation contexts and the first 2,500 contexts of the Hsiao and Schneider corpus were then annotated by two native English speakers and professional translators. They were paid to identify contexts with a critical citation according to our definition.

QUALITY CONTROL WITH THE RE-ANNOTATION OF THE WHOLE CORPUS

As a reminder, the three different methods described in the previous section aimed at providing us with citation contexts which were likely to contain a critical citation, based on the criteria we had chosen. Therefore, we reviewed all the contexts available to us to check that they met our definition and we eliminated those that did not, as well as the duplicates (mainly originating from the ACL corpus, used for four out of the six corpora described in Step 1). During this verification task, we also cleaned up the contexts (e.g. by correcting missing or unnecessary spaces, or by removing XML tags).

(3) DATASET DESCRIPTION

REPOSITORY NAME

Zenodo

OBJECT NAME

20220206_CORPUS_critical_citations_DATA_PAPER.csv

FORMAT NAMES AND VERSIONS

CSV

CREATION DATES

Between 2020-03-01 to 2022-02-06

DATASET CREATORS

Philippe Gambette and Frédérique Bordignon

LANGUAGE

English

LICENSE

CC BY-SA

PUBLICATION DATE

2024-02-22

(4) REUSE POTENTIAL

The corpus can be used as a training set for the automatic detection of critical citations, a functionality that would be helpful in identifying errors (including those resulting from misconduct) or controversies (claims and counterclaims) in the literature, as the online tool Scite_[1] is already offering (but not for free and not with open data). From the point of view of sociologists of science, potential users of this kind of tool, it can be a useful tool to study disagreement in science and the mechanisms of correction of science (in particular thanks to the identifier of the citing paper, whose descriptive metadata can easily be retrieved). It can also be useful for linguists interested in the way criticism is performed in scholarly communication,

1 https://scite.ai/ (last accessed 04/06/2024).

and can shed light on works in the field of English for Specific Purposes (ESP), as much as for language acquisition as for discourse analysis. However, even if this corpus is valuable insofar as it compiles instances of a rare phenomenon, it remains small and will need to be further developed.

ACKNOWLEDGEMENTS

We would like to thank the members of the Cita&Re and NanoBubbles projects, including interns, in particular Karla Avanço who worked on gathering and comparing several corpora of critical citations, and the translators who performed the annotation task.

FUNDING INFORMATION

This work has been initially funded by the French government under the management of the Agence Nationale de la Recherche (ANR-16-IDEX-0003-ISITE FUTURE) and then by the Horizon 2020 Framework Programme of the European Union (H2020-ERC-2020-SyG – Grant Agreement No. 951393).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Data curation: F.B. and P.G.; Methodology: F.B. and P.G.; Software: P.G.; Validation: F.B. and P.G.; Writing – original draft: F.B.; Writing – review & editing: P.G.

AUTHOR AFFILIATIONS

Frédérique Bordignon orcid.org/0000-0002-4918-9137
Ecole des Ponts, Marne-la-Vallée, France; LISIS, INRAE, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France

Philippe Gambette orcid.org/0000-0001-7062-0262
LIGM, Univ Gustave Eiffel, CNRS, Ecole des Ponts, Marne-la-Vallée, France

REFERENCES

Athar, A. (2011). Sentiment Analysis of Citations using Sentence Structure-Based Features. *Proceedings of the ACL 2011 Student Session,* 81–87, P11–3015. https://www.aclweb.org/anthology/P11-3015

Bordignon, F. (2021). A dataset of critical citations contexts. *Mendeley Data, V1*. DOI: https://doi.org/10.17632/2v5d3bpydb.1

Bordignon, F. (2022). Critical citations in knowledge construction and citation analysis: From paradox to definition. *Scientometrics*, *127*, 959–972. DOI: https://doi.org/10.1007/s11192-021-04226-0

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science, 40*(4), 284–290. DOI: https://doi.org/10.1002/(SICI)1097-4571(198907)40:4<284::AID-ASI10>3.0.CO;2-Z

Catalini, C., Lacetera, N., & **Oettl, A.** (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences, 112*(45), 13823–13826. DOI: https://doi.org/10.1073/pnas.1502280112

Dong, C., & **Schäfer, U.** (2011). Ensemble-style Self-training on Citation Classification. *Proceedings of 5th International Joint Conference on Natural Language Processing,* 623–631. https://www.aclweb.org/anthology/I11-1070

Gambette, P. (2024). Scraping-CitaRe (v1.0.2). *Zenodo*. DOI: https://doi.org/10.5281/zenodo.10925183

Hernández Álvarez, M. (2015). Concit-corpus: Context citation analysis to learn function, polarity and influence (Doctoral dissertation, Universidad de Alicante, Spain). https://dialnet.unirioja.es/servlet/tesis?codigo=61958

Hsiao, T.-K., & **Schneider, J.** (2021). Dataset for "Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine". University of Illinois. DOI: https://doi.org/10.13012/B2IDB-8255619_V2; https://doi.org/10.1162/qss_a_00155

Jochim, C., & **Schütze, H.** (2012). Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme. *Proceedings of COLING 2012,* 1343–1358. https://aclanthology.org/C12-1082

Lin, C.-S. (2018). An analysis of citation functions in the humanities and social sciences research from the perspective of problematic citation analysis assumptions. *Scientometrics*, *116*(2), 797–813. DOI: https://doi.org/10.1007/s11192-018-2770-2

Moravcsik, M. J., & **Murugesan, P.** (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, *5*(1), 86–92. DOI: https://doi.org/10.1177/030631277500500106

Oppenheim, C., & **Renn, S. P.** (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science, 29*(5), 225–231. DOI: https://doi.org/10.1002/asi.4630290504

Schneider, J., Ye, D., Hill, A. M., & **Whitehorn, A. S.** (2020). Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics*, *125*(3), 2877–2913. DOI: https://doi.org/10.1007/s11192-020-03631-1

Spiegel-Rosing, I. (1977). Science Studies: Bibliometric and Content Analysis. *Social Studies of Science, 7*(1), 97–113. DOI: https://doi.org/10.1177/030631277700700111

Teufel, S., Siddharthan, A., & **Tidhar, D.** (2006). Automatic classification of citation function. *EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* 103–110. DOI: https://doi.org/10.3115/1610075.1610091

Ye, D., Hill, A., Whitehorn (Fulton), A., & **Schneider, J.** (2020). Citation context annotation for new and newly found citations (2006–2019) to retracted paper Matsuyama 2005. University of Illinois. DOI: https://doi.org/10.13012/B2IDB-8150563_V1

TO CITE THIS ARTICLE:
Bordignon, F., & Gambette, P. (2024). A Corpus of Critical Citations Contexts. *Journal of Open Humanities Data*, 10: 39, pp. 1–6. DOI: https://doi.org/10.5334/johd.215

Submitted: 08 April 2024
Accepted: 28 May 2024
Published: 12 June 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.