Journal of Open Humanities Data

Absorption in Online Reviews of Books: Presenting the English-Language AbsORB Metadata Corpus and Annotation Guidelines

DATA PAPER

MONIEK KUIJPERS
PIROSKA LENDVAI
MASSIMO LUSETTI
SIMONE REBORA
LINA RUH
JONATHAN TADRES
TINA TERNES
JOHANNA VOGELSANGER

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

This paper presents an annotated metadata corpus of English language book reviews from Goodreads and annotation guidelines developed to tag online book reviews for mentions of story world absorption. The metadata corpus includes the segments of each review that have been annotated, the annotation category, the title and author of the book that is reviewed, the rating, the genre of the book reviewed, the length of the review in characters and tokens, and the on- and offset of the annotation. The corpus and guidelines could be used to further investigate the experience of absorption during reading.

CORRESPONDING AUTHOR:
Moniek M. Kuijpers
Digital Humanities Lab,
University of Basel, CH
moniek.kuijpers@unibas.ch

KEYWORDS:
story world absorption; manual annotation; online book reviews; metadata corpus

TO CITE THIS ARTICLE:
Kuijpers, M., Lendvai, P., Lusetti, M., Rebora, S., Ruh, L., Tadres, J., Ternes, T., & Vogelsanger, J. (2023). Absorption in Online Reviews of Books: Presenting the English-Language AbsORB Metadata Corpus and Annotation Guidelines. *Journal of Open Humanities Data*, 9: 13, pp. 1–7. DOI: https://doi.org/10.5334/johd.116

(1) OVERVIEW

REPOSITORY LOCATION

Open Science Framework. "Absorption in Online Book Reviews", https://osf.io/kr4v6/ (DOI: 10.17605/OSF.IO/KR4V6).

CONTEXT

Online book reviews are a relatively new form of reader testimonials that researchers from different disciplines can use to investigate reading experience and evaluation. The research on online book reviews has remained largely theoretical (e.g., Boot, 2011; Murray, 2018; Nakamura, 2013; Rehfeldt, 2017), or when empirical methods were employed, they involved bottom-up, data-driven approaches (e.g., Nuttall, 2017; Nuttall & Harrison, 2020), or a top-down quantitative approach (Milota, 2014) to answer research questions about the reception of *one particular* text. The main aims of the present project were to validate the Story World Absorption Scale (SWAS; Kuijpers, Hakemulder, Tan & Doicaru, 2014) – a self-report instrument to capture experiences of absorption during literary reading – against unprompted reader testimonials found on Goodreads and to build a corpus of online book reviews that could be used for meaningful corpus linguistic as well as qualitative analyses of reader responses, emphasising absorption experience (cf. Kuijpers, Lusetti, Lendvai & Rebora, under review).

(2) METHOD

BUILDING THE CORPUS

We scraped approximately six million English language reviews of nine different genres (i.e., fantasy, romance, thriller, horror, mystery, science fiction, historical fiction, contemporary, classics) from the Goodreads website between the spring of 2018 and the spring of 2019. We selected a subset of reviews to annotate from this larger corpus, taking into account text length and reviewer rating. We eliminated reviews with GIFs and used a word list based on our annotation guidelines to select reviews with a high absorption potential (for more information, see Lendvai et al., 2020; Rebora, Kuijpers & Lendvai, 2020). We then manually annotated this subset of reviews with a group of five annotators using our guidelines developed for this project. Our final corpus consists of 493 curated reviews. After annotation and curation, we added the following metadata to our final corpus: title of the book, author of the book, genre of the book (as voted for by Goodreads members), text length of the review in characters and tokens, annotated segment, annotation categories (book-specific mention of absorption versus mention of absorption in general reading behavior; presence or negation of absorption; absorption dimensions; specific absorption categories), off- and onset of the annotated segment, and annotation round. We have added a ReadMe file to the OSF page where researchers can obtain more information about each of the metadata variables included in the dataset.

Due to copyright and data privacy restrictions to the type of data we are working with, we decided to prepare two versions of our corpus: one with metadata, but without the full review texts (this version is available under the OSF-link presented in this paper). The other version does include the full text reviews, which have been fully anonymized. Researchers who are interested in working with this version of the corpus can contact the first author of this paper to obtain access to the complete corpus for research purposes only. We followed the APA guidelines on protected access open data in this decision (APA, 2023).

DEVELOPING THE ANNOTATION GUIDELINES

The annotation guidelines were developed throughout the annotation process which started in March 2019 and was divided into 15 rounds, the last of which was completed in October 2020 (for a thorough description of the annotation process, see Kuijpers et al., under review). The 18 statements on the Story World Absorption Scale were taken as the point of departure for the annotation guidelines. Throughout the annotation process, we simplified the language used in these statements to better match the language use of Goodreads reviewers and we added 17 annotation categories to the tagset, partly based on further research (i.e., Bálint et al., 2016)

and partly data-driven by what we found in the reviews. To complete the guidelines, we added examples from the reviews for all of the different absorption categories when we could find them. These examples can help other researchers who want to use the tagset to familiarise themselves with the idiosyncratic language found on digital social reading platforms (cf. Nuttal & Harrison, 2020; Pianzola, 2021). Table 1 shows all of the annotation categories and how many times they were used to annotate segments of reviews. The guidelines also contain information about how to assign all aspects of a given annotation category, such as whether a category is present or negated, whether it was an instance of book-specific absorption or a description of how a reader usually experiences absorption, regardless of the book reviewed. Furthermore, the guidelines include comments about the differences between categories that are closely related to one another.

ABSORPTION DIMENSION	ABSORPTION CATEGORY	NUMBER OF TOTAL ANNOTATIONS	ABSORPTION PRESENT	ABSORPTION NEGATED
Attention	A1 (Altered sense of time): *While reading time moved differently*	3	3	0
	A2 (Concentration): *My attention was focused on the book*	12	10	2
	A3 (General sense of absorption): *I was absorbed in the book*	194	186	8
	A4 (No distractions): *I was not distracted while reading*	5	3	2
	A5 (Forgetting surroundings): *While reading I forgot the world around me*	20	16	0
	A6 (Anticipation): *I was on the edge of my seat/I wanted to know what would happen next*	111	108	3
	A7 (Inability to stop reading): *I did not want to put the book down/ I could not put the book down*	150	145	5
Emotional Engagement	EE1 (Perspective taking): *I could imagine what it must be like to be this character*	35	35	0
	EE2 (Sympathy): *I sympathized with this character*	57	53	4
	EE3 (Emotional connection): *I felt a connection to this character*	79	69	10
	EE4 (Empathy): *I felt how this character was feeling*	73	72	1
	EE5 (Compassion for story events): *I felt for what happened in the story*	79	79	0
	EE6 (Anger): *I felt angry at this character*	20	19	1
	EE7 (Fear): *I felt scared for this character*	5	5	0
	EE8 (Emotional familiarity): *I felt like I knew this character*	11	11	0
	EE9 (Wishful identification): *I wish I could be more like this character*	8	8	0
	EE10 (Emotional understanding): *I understood why this character did this*	31	26	5
	EE11 (Parasocial response): *I want to have some kind of relationship with this character*	79	79	0
	EE12 (Participatory response): *I wanted to involve myself in the story world events*	42	42	0
Mental Imagery	MS1 (Imagery of character): *I could imagine what the characters looked/smelled/felt/sounded like*	18	15	3
	MS2 (Imagery of story events): *I could see/hear/feel/smell the story events clearly in my mind*	20	20	0
	MS3 (Imagery of story world): *I could imagine what the story world looked/smelled/felt/sounded like*	27	25	2
	MS4 (Realness): *The character/story world felt real to me*	73	73	0
Transportation	T1 (Presence): *While reading this I was in the story world*	13	13	0
	T2 (Merge of fiction in reality): *Elements from the story world came into my world*	14	14	0
	T3 (Proximity of story world): *The story world felt close to me*	4	4	0

(Contd.)

ABSORPTION DIMENSION	ABSORPTION CATEGORY	NUMBER OF TOTAL ANNOTATIONS	ABSORPTION PRESENT	ABSORPTION NEGATED
	T4 (Deictic shift): *I felt transported to the story world*	19	19	0
	T5 (Part of the story world): *I felt part of the story world*	34	34	0
	T6 (Return deictic shift): *I returned from a trip to the story world*	3	3	0
	T7 (Travel in story world): *I lost myself in the story world/I traveled with the characters through the story world*	26	26	0
Impact	IM1 (Effortless engagement): *It was an easy read/I devoured this book*	108	68	40
	IM2 (Wish to reread): *I will/have reread this book/parts of this book*	116	112	4
	IM3 (Anticipation book series): *I cannot wait to see how this unfolds in the next book*	170	167	3
	IM4 (Addiction): *I am addicted to this book/I cannot get enough of this book*	91	89	2
	IM5 (Lingering story feelings): *The book left me feeling .../This book stayed with me for a while*	194	192	2

Table 1 Annotation layers and categories with number of annotations per category (presence or negation of the category) for rounds 7 to 14 of the annotation process.

Note. Items highlighted in dark grey are more succinct phrasings of the original SWAS statements; medium grey items were taken from the absorption inventory by Bálint et al. (2016); light grey items are additions from the annotation team based on what we found in the reviews.

Note. This table is a reproduction of a table in Kuijpers, Lusetti, Lendvai & Rebora, under review.

(3) DATASET DESCRIPTION

OBJECT NAME

The AbsORB (Absorption in Online Reviews of Books) Corpus and Annotation Guidelines.

FORMAT NAMES AND VERSIONS

Corpus: .csv and .xlsx

Guidelines: .pdf

CREATION DATES

2018-12-01 — 2023-06-15

DATASET CREATORS

Moniek Kuijpers, University of Basel

Simone Rebora, Johannes Gutenberg University Mainz

Piroska Lendvai, Bavarian Academy for Sciences and Humanities

Massimo Lusetti, University of Basel

Lina Ruh, University of Basel

Lukas Renner, University of Basel

Jonathan Tadres, University of Basel

Johanna Vogelsanger, University of Basel

Tina Ternes, University of Basel, and Johannes Gutenberg University Mainz

LANGUAGE

Data and metadata: English

LICENCE

REPOSITORY NAME

Open Science Framework

PUBLICATION DATE

2023-06-01

(4) REUSE POTENTIAL

There are many avenues researchers can take in terms of future research. The corpus can be expanded upon with more reviews from different platforms to see whether language use is different from one online community to the next, or with more metadata, such as the number of responses to a review or the number of times a review is read, which would allow for analyses focused on the social aspects of online book reviews. Another expansion could lie in adding reviews with low ratings, to enable analyses on differences in absorption experiences between low-rated and high-rated books. Relatedly, one could look into the annotations per genre and whether readers of different genres mention different absorption dimensions in their reviews. One such study, for which we provided access to the full-text corpus, used network analysis and found preliminary results that point to a strong similarity in vocabulary within romance reviews and within mystery reviews, suggesting that the reading experience of these groups of readers follow a more stable pattern, compared to other genres (i.e., fantasy, science fiction, horror/thriller) where no strong genre clusters were found (Ternes & Kuijpers, in preparation).

When it comes to the guidelines, other avenues may be explored. For example, the guidelines could be used to annotate a set of different reviews focusing on specific books or genres or they could be used on reviews from different platforms, or even different types of reader responses, such as open survey questions or interview transcripts. Currently a pilot study is being conducted in which the co-occurrence of absorption and changes in the self-concept in reviews on climate fiction is investigated (Loi, Lusetti & Kuijpers, in preparation). Another avenue that is currently being explored is the translation and development of a German language corpus and guidelines in order to investigate whether certain absorption experiences can be, culturally and linguistically, translated to a different language community (Kuijpers, Lusetti, Ruh, Ternes & Vogelsanger, in preparation).

ACKNOWLEDGEMENTS

We would like to thank Lukas Renner, who was part of the original annotator team.

FUNDING INFORMATION

This research was funded by the Swiss National Science Foundation's Digital Lives scheme (Grant project 10DL15_183194) and the Swiss National Science Foundation's Eccellenza scheme (Grant project PCEFP1_203293).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Moniek Kuijpers: funding acquisition, supervision, conceptualization, data curation, writing/editing/review, project administration, methodology, resources, validation

Piroska Lendvai: funding acquisition, supervision, methodology, software

Massimo Lusetti: data visualisation, methodology, investigation, formal analysis, data curation, software, validation

Simone Rebora: funding acquisition, supervision, methodology, investigation, software

Lina Ruh: methodology, writing/editing/review, formal analysis, data curation, validation

Jonathan Tadres: methodology, formal analysis

Tina Ternes: methodology, data visualisation, data curation, validation

Johanna Vogelsanger: methodology, formal analysis, validation

AUTHOR AFFILIATIONS

Moniek Kuijpers orcid.org/0000-0002-3676-5879
Digital Humanities Lab, University of Basel, CH

Piroska Lendvai
Digital Humanities Lab, University of Basel, CH; Bavarian Academy of Sciences and Humanities, Munich, DE

Massimo Lusetti orcid.org/0009-0007-3673-1226
Digital Humanities Lab, University of Basel, CH

Simone Rebora orcid.org/0000-0002-1501-3774
Department of Book Studies, Johannes Gutenberg University Mainz, DE

Lina Ruh
Digital Humanities Lab, University of Basel, CH

Jonathan Tadres orcid.org/0009-0009-5482-8990
Digital Humanities Lab, University of Basel, CH

Tina Ternes
Digital Humanities Lab, University of Basel, CH; Department of Digital Methods in the Humanities and Cultural Studies, Johannes Gutenberg University & University of Applied Sciences, Mainz, DE

Johanna Vogelsanger orcid.org/0000-0001-7919-1222
Digital Humanities Lab, University of Basel, CH

REFERENCES

APA (American Psychological Association). (2023). *Open Science Badges*. Retrieved from https://www.apa.org/pubs/journals/resources/open-science-badges (last accessed: 17.08.2023).

Bálint, K., Hakemulder, F., Kuijpers, M., Doicaru, M., & **Tan, E. S.** (2016). Reconceptualizing foregrounding: Identifying response strategies to deviation in absorbing narratives. *Scientific Study of Literature*, *6*(2), 176–207. DOI: https://doi.org/10.1075/ssol.6.2.02bal

Boot, P. (2011). Towards a genre analysis of online book discussion: Socializing, participation and publication in the Dutch booksphere. *Selected Papers of Internet Research, 12*, 1–16. DOI: https://doi.org/10.5210/spir.v1i0.9076

Kuijpers, M. M., Hakemulder, F., Tan, E. S., & **Doicaru, M. M.** (2014). Exploring absorbing reading experiences: Developing and validating a self-report measure of story world absorption. *Scientific Study of Literature*, 4(1), 89–122. DOI: https://doi.org/10.1075/ssol.4.1.05kui

Kuijpers, M. M., Lusetti, M., Lendvai, P., & **Rebora, S.** (under review). Annotating for story world absorption in online book reviews. *Journal of Cultural Analytics*.

Kuijpers, M. M., Lusetti, M., Ruh, L., Ternes, T., & **Vogelsanger, J.** (in preparation). Absorption in Online Reviews of Books. Presenting the German-Language AbsORB Metadata Corpus and Annotation Guidelines.

Kuijpers, M. M., Rebora, S., Lendvai, P., Lusetti, M., Ruh, L., Vogelsanger, J., & **Ternes, T.** (2023, June 27). Absorption in Online Book Reviews. Retrieved from osf.io/kr4v6 (last accessed: 17.08.2023).

Lendvai, P., Darányi, S., Geng, C., Kuijpers, M. M., Lopez de Lacalle, O., Mensonides, J.-C., Rebora, S., & **Reichel, U.** (2020). Detection of Reading Absorption in User-Generated Book Reviews: Resources Creation and Evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4835–4841. Marseille, France: European Language Resources Association.

Loi, C., Lusetti, M., & **Kuijpers, M. M.** (in preparation). Investigating the impact of reading climate fiction. A Case Study in Empirical Literary Studies Using Online Book Reviews. In J. Alber & R. Schneider (Eds.), *Routledge Companion to Literature and Cognitive Studies*.

Milota, M. (2014). From "compelling and mystical" to "makes you want to commit suicide": Quantifying the spectrum of online reader responses. *Scientific Study of Literature*, 4(2), 178–95. DOI: https://doi.org/10.1075/ssol.4.2.03mil

Murray, S. (2018). Reading online: Updating the state of the discipline. *Book History, 21*(1), 370–396. DOI: https://doi.org/10.1353/bh.2018.0012

Nakamura, L. (2013). "Words with Friends": Socially networked reading on Goodreads. *PMLA, 128*(1), 238–243. DOI: https://doi.org/10.1632/pmla.2013.128.1.238

Nuttall, L. (2017). Online readers between the camps: A text world theory analysis of ethical positioning in *We Need to Talk About Kevin*. *Language and Literature, 26*(2), 153–171. DOI: https://doi.org/10.1177/0963947017704730

Nuttall, L., & **Harrison, C.** (2020). Wolfing down the Twilight Series: Metaphors for Reading in Online Reviews. In H. Ringrow & S. Pihlaja (Eds.), *Contemporary Media Stylistics* (pp. 35–60). London: Bloomsbury Publishing. DOI: https://doi.org/10.5040/9781350064119.0007

Pianzola, F. (2021). Digital Social Reading: Sharing Fiction in the 21st Century. *Work In Progress MIT*. DOI: https://doi.org/10.1162/ba67f642.a0d97dee

Rebora, S., Kuijpers, M. M., & **Lendvai, P.** (2020). Mining Goodreads. A Digital Humanities project for the Study of Reading Absorption. In *Sharing the Experience: Workflows for the Digital Humanities. Proceedings of the DARIAH-CH Workshop 2019*. DARIAH-CH, Neuchâtel. https://zenodo.org/record/3897251

Rehfeldt, M. (2017). Leserrezensionen als Rezeptionsdokumente. Zum nutzen nicht-professioneller Literaturkritiken für die Literaturwissenschaft. In A. Bartl & M. Behmer (Eds.), *Die Rezension. Aktuelle Tendenzen der Literaturkritik* (pp. 275–289). Würzburg: Königshausen & Neumann.

Ternes, T., & **Kuijpers, M. M.** (in preparation). Using networks to navigate a corpus of reader-reviews: An explorative study of absorption expressions in different genres. *Participations. Journal of Audience and Reception Studies.*

TO CITE THIS ARTICLE:
Kuijpers, M., Lendvai, P., Lusetti, M., Rebora, S., Ruh, L., Tadres, J., Ternes, T., & Vogelsanger, J. (2023). Absorption in Online Reviews of Books: Presenting the English-Language AbsORB Metadata Corpus and Annotation Guidelines. *Journal of Open Humanities Data*, 9: 13, pp. 1–7. DOI: https://doi.org/10.5334/johd.116

Submitted: 27 June 2023
Accepted: 08 August 2023
Published: 13 September 2023

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