

Online Book Reviews and the Computational Modelling of Reading Impact

Marijn Koolen^a, Peter Boot^b and Joris J. van Zundert^b

^a*Humanities Cluster - Royal Netherlands Academy of Arts and Sciences*

^b*Huygens Institute for the History of the Netherlands - Royal Netherlands Academy of Arts and Sciences*

Abstract

In online book reviews readers often describe their reading experience and the impression that a book left. The great volume of online reviews makes these reviews a great source for investigating the impact books have on readers. Recently, a reading impact model was introduced that can be used to automatically identify expressions of reading impact in Dutch reviews and that is able to categorise them according to emotional impact, aesthetic or narrative feeling, or feelings of reflection. This paper provides an analysis of the characteristics of the book review domain that affect how this computational model identifies impact. We look at features like the length of reviews, the nature of the website on which the review was published, the genre of book and the characteristics of the reviewer. The findings in this paper provide insight in how different selection criteria for reviews can be used to study various aspects of reading impact.

Keywords

Digital Literary Studies, Reading Impact, Online Book Reviews, Dataset Characteristics

1. Introduction

Online book reviews written by ordinary readers are an important feature of the 'Digital Literary Sphere' [25]. Apart from their obvious commercial interest [6], these reviews also constitute important evidence for how readers read books. Readers often describe their reading experience and the great volume of online reviews makes them a great source for investigating the impact books have on readers [37], as demonstrated by several systematic studies into reading experiences based on online reviews [7, 11, 28]. For an overview see [36].

Recently, we introduced a reading impact model that can be used to automatically identify expressions of reading impact in Dutch reviews and that is able to categorise them according to emotional impact, aesthetic or narrative feeling, or remarks on reflection [4]. The model consists of over 250 rules identifying impact terms or phrases and terms revealing contextual aspects of books, and has been validated against human judgements. These rules can be applied to individual sentences from book reviews, which results in a set of matches. For instance, a sentence containing the word 'meeslepend' (English: 'absorbing' or 'engrossing') expresses narrative impact, but only if the sentence also contains words referring to the book or the story. The model allows us to analyse reading impact at scale. For the first time, an

CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands

✉ marijn.koolen@gmail.com (M. Koolen); peter.boot@huygens.knaw.nl (P. Boot); joris.van.zundert@huygens.knaw.nl (J.J.v. Zundert)

🌐 <https://marijnkoolen.com/> (M. Koolen)

🆔 0000-0002-0301-2029 (M. Koolen); 0000-0002-7399-3539 (P. Boot); 0000-0003-3862-7602 (J.J.v. Zundert)

© 2020 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

arbitrarily large amount of book reviews can be computationally analysed to investigate how individual books, or entire genres, affect their readers.

A pertinent question is how exactly tens of thousands of reviews can be harnessed to gauge the impact of a novel on its audience. This cannot be a simple matter of tallying rule matches, because some books have thousands of reviews while most others have none or only a few, and some reviewers produce hundreds of reviews while many others write a few. Similarly, how do we compare and aggregate a score of seemingly “calm and collected” reviews with a small number of extremely passionately enthusiastic reviews that relate to the same work? Simple tallying might justifiably lead to concerns about reductive handling of data, which is a concern that is commonly raised as a problem with respect to quantified and computational approaches in digital humanities [13, 12]. To answer this question we need to improve our understanding of how the varying make up of reviews and how different selections, categorizations, and aggregations of rule matches affect the analytical outcome of the model.

This paper provides a first analysis of the characteristics of the book review domain that affect how this computational model identifies impact. We look at features like the length of reviews, the nature of the website on which the review was published, book genre and the characteristics of the reviewer. The operationalisation of our model prompts a number of research questions:

- How can we translate matches of individual impact rules into an overall impact score for a review or a set of reviews for the same book?
- How are impact rule matches related to other review characteristics, such as the length of the review or the website for which the review was written?
- How are impact rule matches related to reviewers, reviewed books and book genres?

We first discuss the background of analyzing reading impact and book reviews in Section 2, then describe the characteristics of the review dataset we use in Section 3. Then we analyse the relationship between review characteristics and impact matches and how to aggregate these into interpretable impact scores in Section 4. We close this paper with a discussion of our findings, their implications for future research and the limitations of our work in Section 5.

2. Online Reviews and Reading impact

In this section we discuss related work on online book reviews before describing the Reading Impact Model in more detail.

2.1. Research on Reading Impact

The impact that reading fiction has on a reader has been studied for several decades. So far, this has mostly been done through interviewing readers [39, 38], studies with reader responses to short stories and passages [27, 24, 21, 20], or through theoretical argument [29]. Most of the effects that were found centre on emotions, e.g. enjoyment, empathy [19], sympathy and aesthetic response [24]. Readers also report effects of personal transformation [24, 39, 38], self-reflection [21, 20] and changing beliefs about the real world [15]. The reading impact model of Boot and Koolen [4] uses the four categories of Koopman and Hakemulder [21], namely, *general emotional impact*, *narrative feeling*, *aesthetic feeling* and *reflection*.

2.2. Research on Online Book Reviews

Online book reviews have also been used as a source to study literary reception, mostly through close reading of a relatively small number of reviews [16, 14, 26, 48]. Spiteri and Pecoskie [42] analysed 536 online book reviews to derive a taxonomy of reader’s experiences. Driscoll and Rehberg Sedo [11] analysed language use in 692 Goodreads reviews using feminist theory to understand how reviewers articulate intimate reading experiences. There are also some computational studies using large scale datasets, e.g. Hajibayova [17] looked at language use in 475,000 Goodreads reviews to investigate reader’s perceptions and behaviours. Thelwall [46] looked at author gender preferences of readers in 200,000 Goodreads reviews. Finally, Rebora et al. [35] used textual entailment and text reuse detection methods to classify 3,500 sentences from Goodreads reviews to the Story World Absorption Scale by Kuijpers et al. [22]. Lendvai et al. [23] recently released a corpus of these sentences, manually annotated using the absorption scale.

Some potential issues with insincere reviews have been signaled. Authors and publishers may game the system by writing positive reviews of their own books and negative reviews of competitors’ books [41]. Reviews can also be bought, with companies offering to write reviews for profit [44]. There are some characteristics of insincere reviews that can be used to (semi-) automatically detect them with some level of reliability [40]. However, there is also a gray area of reviews for which it is nigh impossible to judge their sincerity. Ott et al. [31] developed a generative model for deception along with a deception classifier to estimate the prevalence of fake reviews for hotels from six websites and found that 2-6% of reviews are likely deceptive.

Reviews can also be written as a form of identity formation: reviewers are not just focusing on their actual reading experience but care about how they are perceived by others and may report a reading experience that is partly informed by their desired outcome [8, 32]. Thelwall and Kousha [47] found that the book-based social network Goodreads is a genuine hybrid platform in which the majority of users engage with both book-based activities (adding, rating and reviewing books they have read) and social activities (building a network of friends and followers and uploading photos). Beyond these social considerations, reviews are also a genre, with online reviews perhaps developing their own conventions [10, 43, 1, 45].

Finally, differences have been highlighted between reviews on book selling sites and those on social book review sites, especially as to objectives and motivations of reviewers for writing reviews [9]. On book selling sites reviewers write more purchase oriented reviews and sometimes include aspects of the selling process. They are also more likely to provide more extreme values in their ratings and reviews, perhaps to influence potential buyers, which is less directly relevant on platforms where no books are sold.

2.3. The Reading Impact Model

The Reading Impact Model [4] was recently published as a generic model for studying expressions of reading impact in book reviews. The model consists of 257 rules.¹ Each rule contains an impact term that is either a single word, like *adembenend* (English: ‘breathhtaking’), or a phrase, like *op het puntje van (me/mijn/je) stoel* (‘on the edge of (my|your) seat’). For single word terms, the rule is checked against the lemma of a word in the sentence. The phrases are matched against sentences as regular expressions, so no lemma information is used. Instead,

¹The model is available from <https://github.com/marijnkoelen/reading-impact-model>

the phrases contain morphological variants (*me/mijn*) to compensate for such variation in the surface form of the sentence.

The model uses the four impact categories of Koopman and Hakemulder [21]. *Emotional impact* is a generic impact category, while *narrative feeling* or *narrative impact* is impact of narrative aspects like the story, plot or characters. *Aesthetic feeling* or *aesthetic impact* is impact of style. Finally, *reflection* is impact that makes the reader reflect on things external to the book, which could be their own thoughts, memories or attitudes, or ideas about other people or things. The model was validated using a set of sentences annotated by multiple persons, with the rules for *emotional impact*, *narrative feeling* and *aesthetic feeling* corresponding well to human judgements. However, *reflection* is not well captured, which according to Boot and Koolen [4] is probably due to a lack of rules to cover all the ways in which a reviewer can express reflection about the external world.

Several rules have the same impact term, like *schitterend* (English: 'beautiful'). If the sentence contains no specific book aspect, the expression will be categorized as 'emotional impact', a general category with affective terms. However, if the sentence contains both *schitterend* and a story aspect like *verhaal* ('story') or *personage* ('character'), the expression is labeled as 'narrative impact'. If *schitterend* co-occurs with a stylistic aspect like *geschreven* ('written') or *schrijfstijl* ('writing style'), it is labeled as 'aesthetic impact'.

3. Review Characteristics

In this section we look at the characteristics of the reviews in the dataset and how these characteristics are related to matches from the Reading Impact model.

It is possible that certain kinds of reviews or certain kinds of reviewers have more matches with the Reading Impact model than others, which might skew the overall picture we get for the reviews of a certain book, author or genre. As is typical of web data [18, 30, 34], there are various aspects that can lead to skewed distributions. Differences in popularity will result in some books having thousands of reviews while most others will have none or just a few. The overall group of reviewers is huge and widely varied, with some highly prolific reviewers writing hundreds or thousands of reviews, and again most others writing only a single review. Some write very personal or highly idiosyncratic reviews while others write fairly standard or superficial reviews. Many reviews will be short but some will be very long. Long reviews have a higher a priori probability of matching impact rules, as they have more sentences and/or more words per sentence. Short reviews and reviews with short sentences have lower probabilities.

Reviewers on book selling platforms like Amazon have additional aspects to review, such as the acquisition process, and different motivations for writing the review, including to share their experience with and opinion of the book seller [9]. All these different aspect may affect how the reading impact of a book should be adequately pieced together from different reviews.

3.1. Review preprocessing

The impact model comes with a matcher function that accepts sentences either as plain text strings or as syntactically parsed trees in the format created either by Alpino² or spacy.io³. Many rules are based on the lemma of a word instead of specific morphological variants. We

²<http://www.let.rug.nl/vannoord/alp/Alpino/>

³<https://spacy.io>

Table 1

Descriptive statistics of the sources of book reviews

Source	# Reviews	# Sentences		# Words		
		Total	Per review	Total	Per review	Per sentence
Boekmeter	7,250	79,005	10.9	1,642,051	226.5	20.8
Bol	254,081	1,308,628	5.2	21,797,198	85.8	16.7
Dizzie	26,880	244,990	9.1	4,598,881	171.1	18.8
Goodreads	90,602	667,691	7.4	10,068,300	111.1	15.1
Hebban	48,783	726,584	14.9	13,867,809	284.3	19.1
LTL	7,004	63,177	9.0	1,205,802	172.2	19.1
WLJN	38,210	272,646	7.1	4,790,380	125.4	17.6
Total	472,810	3,362,721	7.1	57,970,421	122.6	17.2

preprocessed all reviews using NLTK [2] to split the whole text into sentences, then using Alpino for the syntactic analysis of the individual sentences.

3.2. Review lengths by platform

The collection of book reviews that we use consists of 472,810 reviews of fiction, written in Dutch. The majority of reviews come from an earlier collection created by Boot [3], which we extended with additional reviews from Goodreads. The reviews come from seven different review websites (Table 1):

- Boekmeter:⁴ a Dutch website with over 13,000 members who can rate and review books.
- Bol:⁵ a major Dutch webshop that sells a huge range of products including books. Buyers of products are invited to write a review of their product, so not everyone can review any book they like. Several reviews discuss the selling and shipping process.
- Dizzie:⁶ a Dutch book review platform that is no longer online, on which members could discuss and reviews books.
- Goodreads:⁷ an international social book cataloguing website with over 90 million members and over 90 million reviews, where any member can write a review on any book they choose. Our sample of reviews was crawled targeting Dutch language reviews by focusing on Dutch authors, although the reviews also cover thousands of books by non-Dutch authors. Because of this focus, the set of reviews is most likely not representative of all Dutch reviews and reviewers on Goodreads.
- Hebban:⁸ a Dutch book reviewing platform with over 200,000 members who can review and discuss books they have read or want to read next. Members can review any book that is in the Hebban catalogue, and ask for additional books to be added by the platform editors.

⁴<https://www.boekmeter.nl>

⁵<https://www.bol.com/nl/>

⁶Originally at <https://dizzie.nl>, a short description can be found at <https://mustreads.nl/dizzie-nl/>.

⁷<https://www.goodreads.com>

⁸<https://www.hebban.nl>

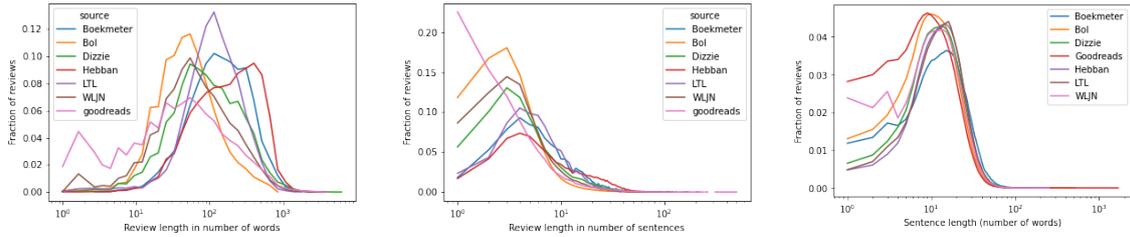


Figure 1: Distribution of reviews over review length in terms of number of words (left), and sentences (middle) and number of words per sentence.

- Lezers Tippen Lezers (LTL):⁹ a Flemish website where readers can find tips on what to read next and post reviews.
- Wat Lees Jij Nu (WLJN):¹⁰ a small Dutch book review website that is no longer online. The site places no restrictions on what members can review.

The reviews from the different platforms have some different characteristics, as shown in Table 1. The majority of reviews come from Bol, which are shorter on average (85.8 words) than those of other platforms. The Dutch Goodreads reviews have an average number of words of 111.8 and an average number of sentences of 7.7, which is somewhat longer than reported for English reviews from Goodreads by Dimitrov et al. [9] (87.8 words and 5.0 sentences). This may be due to general length differences between English and Dutch sentences (although we have not found clear evidence for this, see e.g. [33] for statistics on aligned sentences), or due to differences in how the reviews were collected. Dimitrov et al. [9] focused on reviews for books in the *biography* genre, whereas we started our crawl from a list of Dutch fiction titles.

The distribution of review length for the seven platforms is shown in Figure 1, using the number of words and sentences as units. The Y axis shows the probability of a review having a certain length. This normalisation to probabilities allows us to compare the reviews from different subsets of the collection. The plots use a logarithmic scale on the X axis. For the Y axis, the fraction of reviews are shown on a linear scale. The number of words (left) show that the review lengths for the seven platforms are roughly log-normally distributed.¹¹ This means the difference between using 100 to 250 words is similar to the difference between using 10 to 25 words. Why is the word distribution log-normal? Probably because the review length is positive but open-ended for most platforms. Many reviews have at least a few dozen words to a 100 words (where the peaks of the distributions are). It is possible (though unlikely, as the plots show) to use thousands of words more than the median, but only a few dozens of words less than the median.

The word distributions show significant shifts between the distributions, indicating that there are platform-specific factors playing a role in how much text reviewers write. The reviews on Boekmeter, Hebban and Lezers Tippen Lezers (LTL) have fewer short reviews than the other platforms. Bol deviates strongly at the longer end of the distribution, where its distribution

⁹<http://lezerstippenlezers.be>

¹⁰Originally at <http://www.watleesjij.nu/>, a short description can be found at <https://mustreads.nl/watleesjij-nu/>

¹¹We confirmed this by fitting theoretical models on the data and computing the Residual Sum of Squares for the normal ($RSS = 1.47e^{-5}$), log-normal ($RSS = 2e^{-7}$) and exponential distributions ($RSS = 1.3e^{-6}$).

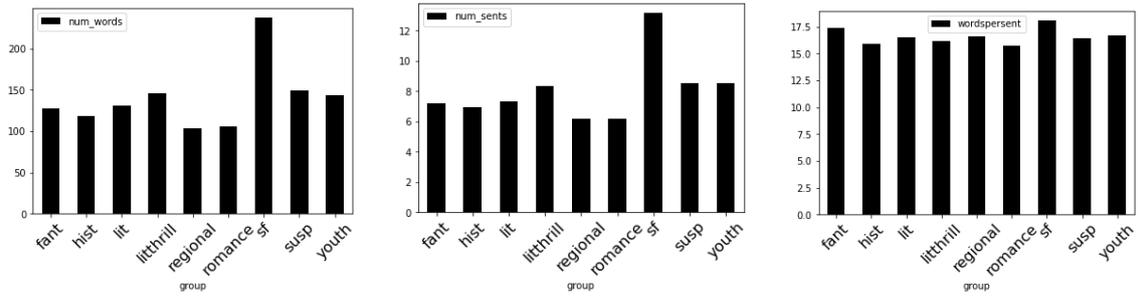


Figure 2: Review lengths by genre in terms of number of words (left), sentences (middle) and words per sentence (right). In a one-way ANOVA the three figures’ p-values are below 0.001.

drops faster than the rest and stops at around 1000 words. This suggests an artificial limit, e.g. it looks like Bol restricts reviews to be at most 4000 characters. The longest review from Bol in our dataset is exactly 4000 characters, with another 178 reviews between 3950 and 4000 characters. Another deviation is seen in the Goodreads set, with many very short reviews of just two words. A manual check reveals that these are typically reviews saying e.g. ‘4 sterren’ (4 stars). Given that on Goodreads users also provide star-based ratings these reviews simply mimic the rating and provide no additional descriptive information. For computing reading impact it would be relatively easy to identify these reviews and, we argue, filter them out without negative consequences for the reading impact analysis. An additional advantage would be that the remaining Goodreads reviews have a length distribution that is more similar to those of the other platforms. Still, when we want to compare reading impact in reviews across platforms, we should take into account these differences in length distribution.

The length distribution by number of sentences is shown on in the middle of Figure 1. There are few reviews longer than 40 sentences, but some are over 200 sentences. The longest is over 500 sentences and contains a very detailed summary of a book on finance.¹² Hebban has a higher proportion of reviews with more than 20 sentences than the other platforms.

Finally, the distribution of number of words per sentence (Figure 1) shows that on some platforms many reviews have some very short sentences, notably Goodreads and WLJN. But all platforms show a peak between 10 and 20 words and dropping off sharply after that, with virtually no sentences over 40 words. Also in terms of sentence length, the reviews from different platforms are comparable.

3.3. Review lengths by genre

As well as by platforms, review lengths differ by other characteristics. In Figure 2 we show the average review length for nine genre groupings:¹³ fantasy, historical novel, literature, literary thriller, regional novel, romance, science fiction, suspense and youth (fiction for children of 13 years and older). As we note, reviews in the science fiction genre are clearly longer than in the other genres. This is mostly due to a larger number of sentences;¹⁴ the number of words per

¹²We filtered our reviews to be only on fiction books, but there are mistakes in the available genre information.

¹³The groups are combinations of on publisher-assigned genre codes (NUR). For instance, literature is a combination of NUR 301 (Dutch literary novel or novella) and 302 (translated literary novel or novella). The figures are based on the 242150 reviews for which we have the books’ NUR code.

¹⁴Post-hoc analysis using the Tukey-HSD test show SF differs significantly from the other genres for number of sentences and number of words.

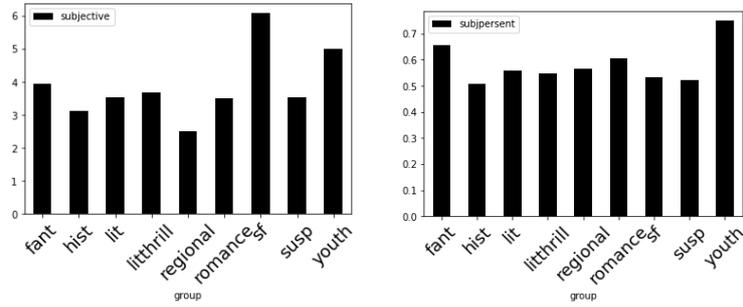


Figure 3: Subjectivity by genre in terms of word count (left) and word count per sentence (right). In a one-way ANOVA the figures' p-values are below 0.001.

sentence vary only slightly.

In the next section (4) we will look at how these lengths affect impact. Here we look at how genre affects subjectivity, which we define as the number of occurrences of first and second person singular pronouns. We assume that sentences where the reviewers refer to themselves (e.g. 'it made me laugh') or to the reader of the review ('it will leave you speechless') are prime candidates to look for expressions of reading impact. We use LIWC to compute these numbers [5]. In Figure 3 we show these counts by genre, as raw numbers as well as per sentence. We notice that in terms of raw numbers, the science fiction genre scores highest. However, when we look at subjectivity per sentence, it becomes clear that science fiction readers use less subjectivity references per sentence than e.g. readers of fantasy. The most subjective readers are fantasy readers and , especially, younger readers.¹⁵ If subjectivity per sentence depends on genre, that might imply that we should not simply 'correct for' the number of sentences when we try to define an impact score in section 4.

3.4. Review lengths by reviewer and per book

The average review length is 7.1 sentences. But the average length of the reviews per book depends very much on the book. Figure 4 shows the distribution of the average number of reviews by book (left) and by reviewer (right). Some titles with on average very short reviews are *Harry Potter and the Half Blood Prince*, *The Devil Wears Prada* and *Shopalicious!* Titles with on average longer reviews include the thrillers *I Am Pilgrim* and *Passenger 23*. Similarly, reviewers vary widely in the amount of text that they produce.

If review length is related to the number of impact expressions, then it is important to understand how review length is related to other aspects of reviews. For instance, popular books have more reviews than unpopular or obscure books, but their reviews might also differ in length. Perhaps popular books get more short reviews than less popular books. The same applies to reviewers. Reviewers who write many reviews may write longer or shorter reviews than reviewers who write only a few reviews. We split the review set into three subsets with a low, medium and high number of reviews per book or per reviewer. Because the distribution is highly skewed, we use thresholds at different orders or magnitude, with low, medium and high respectively corresponding to $0 < x \leq 10$, $10 < x \leq 100$ and $x \geq 100$ reviews per book or per reviewer. Figure 5 shows the distribution of review length for different subsets of reviews, for books (left side) and reviewers (right side). For the split in reviews per book, the Low, Mid

¹⁵Confirmed by post-hoc analysis.

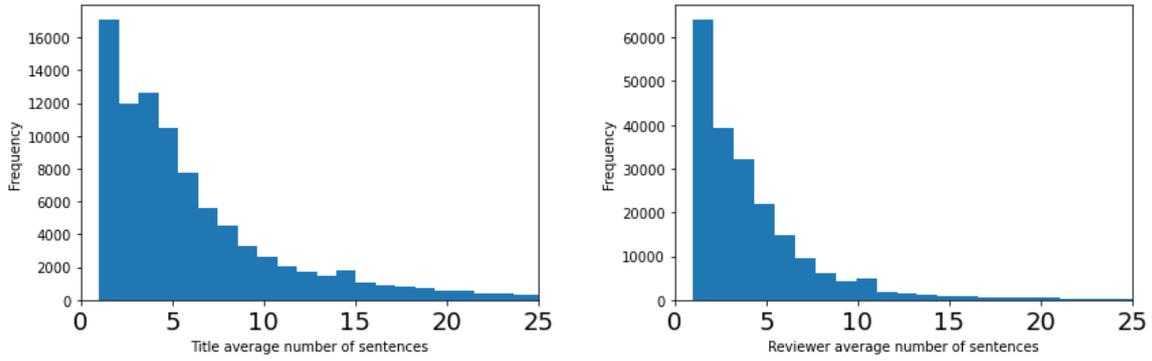


Figure 4: Frequencies of average review length in sentences by book (left) and by reviewer (right), cut off at 25.

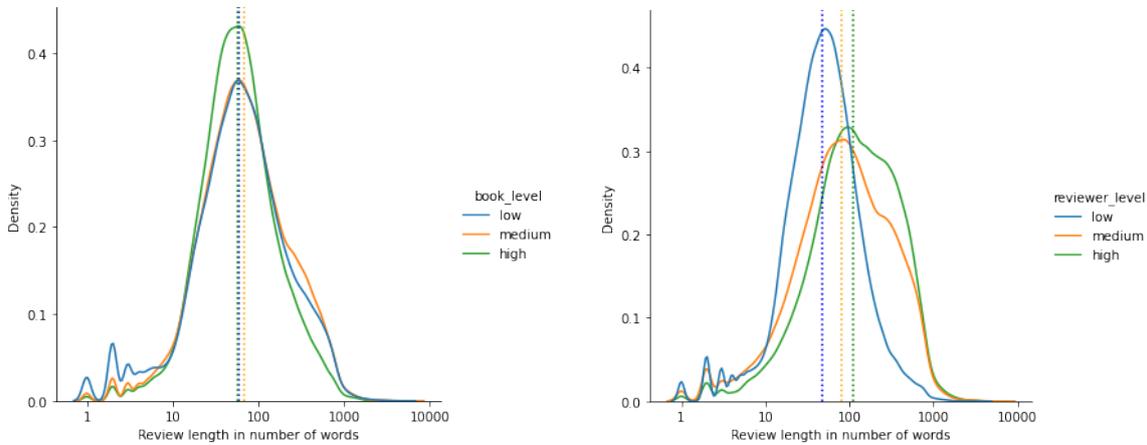


Figure 5: Kernel density estimation of review length (in number of words) for different subsets based on books (left) and reviewers (right) with a low, medium or high number of reviews. The dashed vertical lines represent the mean length using the logarithm of the number of words.

and High frequency sets contain 38%, 47% and 14% of all reviews respectively. For the split in reviews per reviewer, the sets contain 57%, 23% and 19% of the reviews.

On the left side of Figure 5, this frequency split is shown for the number of reviews per book. The different subsets have very similar distributions, with the reviews for high frequency reviewed books having a larger fraction of short reviews and a lower fraction of long reviews. The dotted vertical lines show the per-subset mean length of the logarithm of the number of words. The KL-divergence, which is a number to quantify the distance between two distributions, between the overall distribution and each of the three subsets is 0.01, 0.00 and 0.02 respectively. Although the differences are statistically significant (a one-way ANOVA with Tukey post-hoc tests shows all paired differences are significant with $P < 0.001$), they are very small. The take away message is that, as far as review length is concerned, popularity (in terms of number of reviews per book) causes no big distinction between books with different numbers of reviews. So even though individual reviews differ drastically in length, when aggregated at the book level, book popularity does not introduce a hurdle in comparing reading impact scores between books.

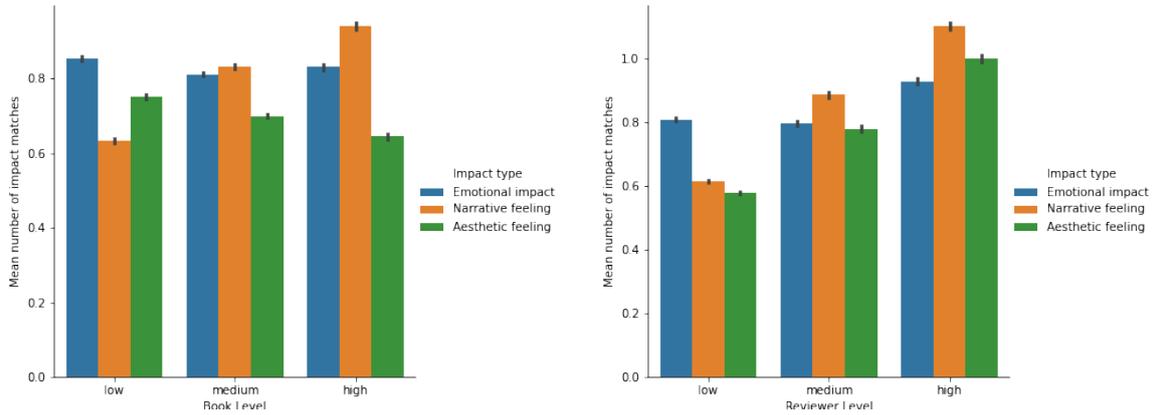


Figure 7: The proportion of reviews that have at least one impact expression for different types of impact, for books (left) and reviewers (right) with a low ($x \leq 10$), medium ($10 < x \leq 100$) or high ($x > 100$) number of reviews.

longer because reviewers include a longer summary of the story, presumably more factual? In that case correction for length of review would be unfair to the verbose reviewer.

A total of 347,491 reviews in the dataset have at least one impact match (73% of all reviews). The probability that a review has at least one impact rule match is shown on the right in Figure 6. We see that, at the level of whole reviews, very short reviews have a very low probability of matching impact rules. One or two word reviews have 2% probability, with three, four or five words this goes up to 5%, 7% and 10% respectively. We draw two lessons from this observation. First, the low probability at the review level is in stark contrast with probability at the sentence level shown on the left of Figure 6, where very short (one or two word) sentences have an almost 20% probability of having an impact match. These very short reviews necessarily have very short sentences, but these are not ones with impact matches. Therefore, the very short sentences with impact matches must come from longer reviews. Second, the low probability at the review level means that in a set of reviews for e.g. the same book, a higher proportion of very short reviews results in a lower proportion of reviews with impact matches. In Figure 5 we saw that popular books have a somewhat higher proportion of very short reviews than less popular books. This finding suggests that, to compare popular and less popular books, the different proportions of short reviews needs to be taken into account. Or more generally, that in weighting the importance of finding an impact match in a review, we should take the review length into account.

4.3. Number of Reviews per Book and Reviewer

Are there important differences in how often reading impact expressions are found in reviews for books with different levels of popularity, or in reviews written by reviewers with different levels of reviewing frequency? These questions are important to understand whether such differences in frequency are an underlying cause of any differences observed when comparing reading impact across a specific set of books. If reviews of popular books would be much more likely to contain, for instance, expressions of aesthetic impact than less popular books, then observing this difference in comparing a Harry Potter book against a relatively unknown fantasy novel does not necessarily tell us much about how these specific books differ in terms

of aesthetic impact. It is possible that popular books are more popular because their aesthetic impact is part of their appeal, but it is also possible that their popularity draws a reviewer's attention to the writing style. The same goes for comparing two books, one of which having reviews by mostly frequent reviewers, the other having reviews by mostly infrequent reviewers. If frequent reviewers write longer reviews and have a checklist of aspects to include in their review, while infrequent reviewers write short reviews with only the first thing that comes to mind, then it is possibly more worthwhile to think about why different books draw different types of reviewers than looking at the impact they express. If we find no frequency effects, it is easier to interpret differences between specific (sets of) books.

The impact of book and reviewer frequency in the collection on the mean number of impact expressions per review of a certain impact type, is shown in Figure 7. We use the same frequency levels as in Section 3.4. On the left the proportions are shown for books with a low, medium and high number of reviews. In the following analysis, we measured statistical significance of differences using the Kruskal-Wallis test by ranks and Bonferroni-Holm post hoc tests. These are non-parametric tests that do not assume normality of data distributions. All differences between impact types per book level are significant ($P < 0.001$), between book levels per impact type we find non-significant differences for Emotional impact between low and high ($P = 0.74$) and for Aesthetic feeling between medium and high ($P = 0.49$). For emotional impact, the differences between books with a low, medium or high number of reviews (the blue bars) is small. For the two more specific impact types, the differences are more pronounced. Low popularity books tend to provoke more expressions of aesthetic feeling than more popular books, but fewer expressions of narrative impact. This could mean that more frequently reviewed books are more narrative-driven and draw the reader into the story world, while less frequently reviewed books tend to have more noticeable writing styles. It could also mean that books with few reviews tend to be read and reviewed by reviewers who focus more on style. On the right the proportions are shown for reviewers, and all differences are significant with $P < 0.001$. Here we see a different pattern. First, reviewers who write more reviews tend to use more expressions of impact of all types. This is likely related to the fact that they tend to write longer reviews. Second, reviewers who write few reviews use more generic expressions of emotional impact than expressions of either aesthetic or narrative feeling, while more prolific reviewers tend to use more expressions of narrative feeling than generic expressions of emotional impact. Third, there is an upwards trend in the relative proportion of aesthetic feeling to emotional impact expressions. As mentioned above, it is possible that reviewers who write many reviews are more likely to go through a list of aspects they want to cover in their reviews, e.g. plot and stylistic elements. This would make sense if reviews follow genre conventions [10], with frequent reviewers possibly being more aware of these conventions or developing their own conventions.

4.4. Correlation of Matches at the Review Level

Obviously we would want to control for possible skewed relations between the number of impact matches of different kinds in a review. The question in this case is: is there some meaningful relation between e.g. the number of matches revealing narrative impact and those matches related to style? We can gauge this by pairwise plotting the number of matches of each impact category per review (Figure 8). The correlations of these pairwise scatter plots is 0.25 on average (cf. the column "Correlation" in table 2). The relatively low correlations may mean that in further analysis it would be advisable to normalize the number of matches in categories

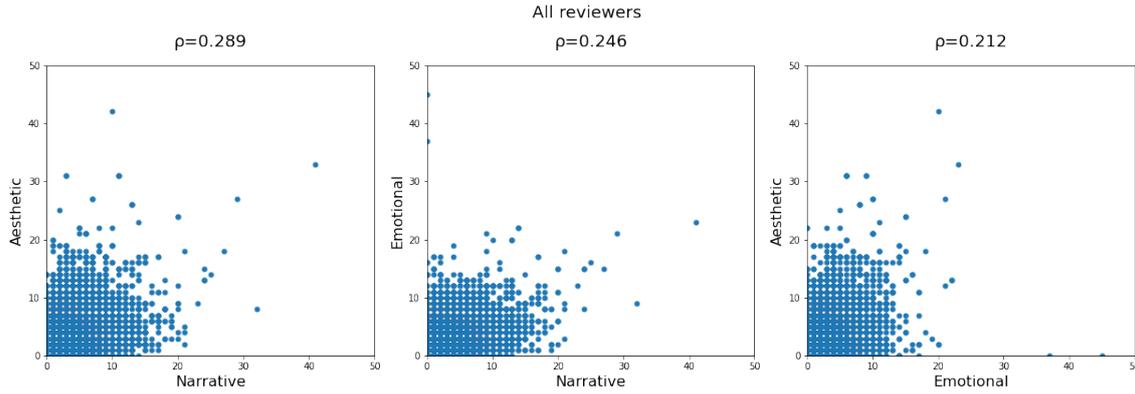


Figure 8: Pair wise scatter plots of the number of matches per impact category in reviews.

Table 2

Pearson's correlation between average counts in the different impact categories

Impact category combination	Correlation	Reviews by reviewer frequency		
		Low $0 < x \leq 10$	Medium $10 < x \leq 100$	High $100 < x$
Narrative \sim Aesthetic	0.29	0.15	0.32	0.35
Narrative \sim Emotional	0.25	0.09	0.36	0.35
Emotional \sim Aesthetic	0.21	0.07	0.30	0.32

when comparing across different impact categories.

We note that these correlations are not evenly distributed across reviewers. This can be shown by dividing reviews again in three categories depending on whether they are written by low, medium, or high frequency reviewers. For this we use the same procedure as in Section 3.4. The number of reviews per reviewer as a distribution is, as mentioned before, very heavily skewed. Reviewers writing only one review account for about 36 percent of the total amount of reviews while a long tail of reviewers produces hundreds of reviews per person with one reviewer topping out at 827 reviews. As can be gauged from table 2 the correlation between numbers of impact matches from different categories climbs as the number of reviews per reviewer increases. This trend becomes the more clear when we refine the procedure for dividing reviews according to reviewers' frequencies of reviewing so we can produce a more continuous graph of correlations (cf. Figure 9). The trend may be an effect of reviewers that write reviews more frequently adopting a more regular structure for reviews, dedicating balanced space to different kinds of reader interests. This effect may thus be indicative of a developing genre convention.

A closer inspection of outliers having particular large and unbalanced impact matches (e.g. a review with 42 aesthetic feeling matches but only 10 narrative feeling matches) reveals that these reviews are almost without exception "user generated data" artefacts that are rather atypical for online reviews. Mostly they are aggregate reviews constructed by compounding the findings of four or five readers into one review. Such reviews should of course not be ignored, but they do not seem to provide much useful additional information as to the point in question of determining how reviews can be made comparable for reader impact.

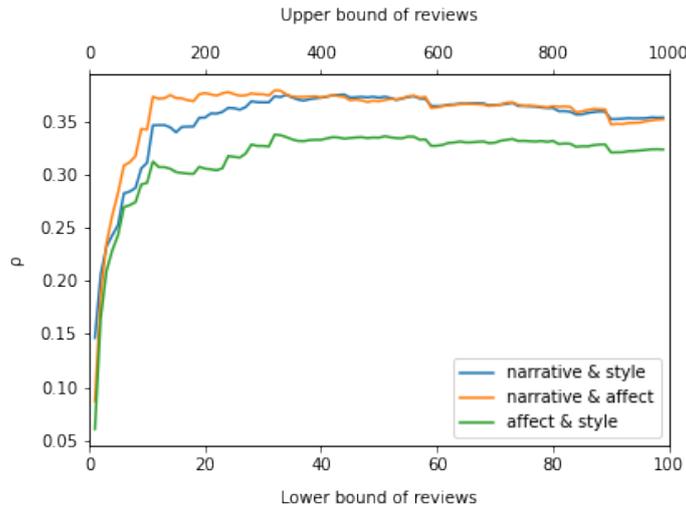


Figure 9: 'Continuous' pairwise correlations when reviews are divided across 100 levels of reviewer review frequency. The x-axes indicate the lower and upper boundary of the review frequencies for which correlations were computed.

4.5. Combining Evidence into a Score

The previous sections have shown that, when comparing reviews or sets of reviews in terms of identified impact expressions, simple counts do not represent a meaningful impact score. There are different characteristics that affect how likely it is to find impact expressions in reviews, such as the length of a review. For very short reviews, it is much less likely than for reviews of a few hundred words. Therefore, to find an expression of impact in a very short review is more surprising – and we assume more significant – than finding one in a long review. This suggests we should weight impact rule matches differently based on review length.

But how can we incorporate these different characteristics as part of the evidence for calculating an impact score?

Starting from intuition, we consider two assumptions. One is that a short review is a signal that not all impact has been expressed. Another is that the book had little impact. A simple solution to compensate for length is dividing the number of impact matches by length. But this takes into account only the first assumption. A single word review with a single impact match would score ten times higher than a 100 word review with 10 impact matches. To allow for the second assumption as well, length normalization should not be linear.

The probability curves for *Emotional impact*, *Narrative feeling*, and *Aesthetic feeling* on the right-hand side of Figure 6 show an almost linear trend for the logarithm of review length (in words). This suggest an alternative solution, namely, to divide by the log of the length. Although the curves drop after 900 words, this is possibly due to the size and composition of the review collection. A larger sample with more reviews from platforms that introduce no length constraints might show curves that flatten out above a certain length instead of drop down, as there is no reason why a longer review cannot have many expressions of reading impact. A generic way of compensating for review length would be to use a logarithm-weighted normalization. That is, the number of matches $I(r_i)$ for a review r_i is weighted by $\frac{1}{\log(|r_i|)}$, where $|r_i|$ is the length of the review.

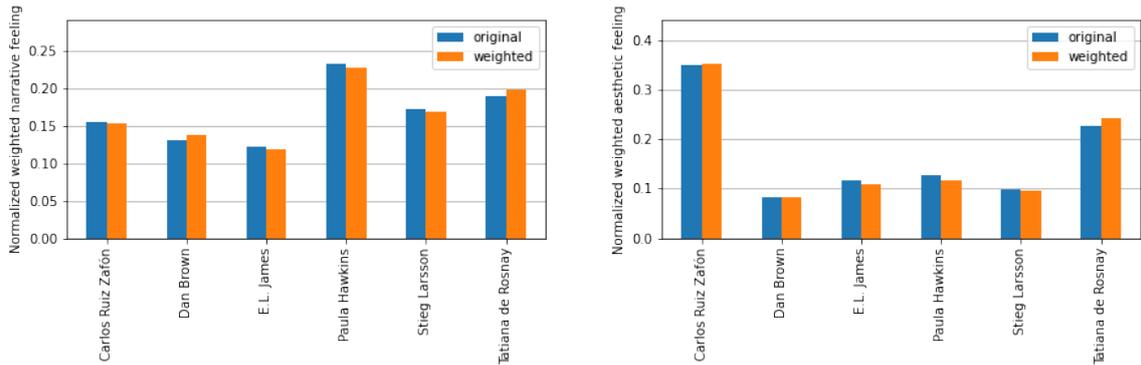


Figure 10: The impact of weighting the number of impact matches by the log-length of reviews for six frequently reviewed books for narrative feeling (left) and aesthetic feeling (right).

We show the impact of length-weighted normalization on the reading impact scores in Figure 10 for narrative feeling (left) and aesthetic feeling (right). The bars show the relative average impact score¹⁶ for six popular books: *The shadow of the wind* by Carlos Ruiz Zafón, *The Da Vinci code* by Dan Brown, *Fifty shades of grey* by E.L. James, *The girl on the train* by Paula Hawkins, *The girl with the dragon tattoo* by Stieg Larsson and *Sarah’s key* by Tatiana de Rosnay. The blue bars show the relative average impact per review using the number of matches, while the orange bars show the weighted scores. The overall differences between the distributions of absolute and normalized number of impact matches are significantly different (Kruskal-Wallis, $P < 0.001$). However, the weighting has almost no effect on the statistical significance of differences between books compared to the original impact score. In most cases, what is significantly different using the impact counts is significantly different after normalization, and similar for what is not. The first thing to note is that by averaging over a large number of reviews (*The girl on the train* has the fewest reviews, with 504 reviews), including many very short reviews, the weighting has a relatively small impact on the relative scores. But there are subtle changes. It is not the case that weights compress all the scores so that the impact always becomes more similar across books. The highest narrative impact score—for Paula Hawkins’ *The girl on the train*—drops while for a few of the others the scores go up, but several lower scoring books always have a lower weighted relative score. The differences in score between *The girl on the train* and the other books are all significant ($P < 0.001$), apart from *Sarah’s key* by Tatiana de Rosnay. After weight normalization, E.L. James scores significantly different on narrative feeling than all the others. For aesthetic feeling (on the right) we see a similar pattern. The weighting does not necessarily reduce the differences between books. Here, the scores for the books by Carlos Ruiz Zafón and Tatiana de Rosnay are significantly different from each other and all the other books (Conover-Iman, $P < 0.001$), the others are not different from each other ($P > 0.05$).

However, if we compare the weighted and non-weighted relative average impact scores for an individual book across platforms, the typical pattern is that the difference between the platform with the highest average score (Hebban, which has reviews that tend to be longer

¹⁶The relative score of each book is the average impact score for that book divided by the sum of averages of all six books. By turning both the weighted and non-weighted scores into proportions, we can directly compare them.

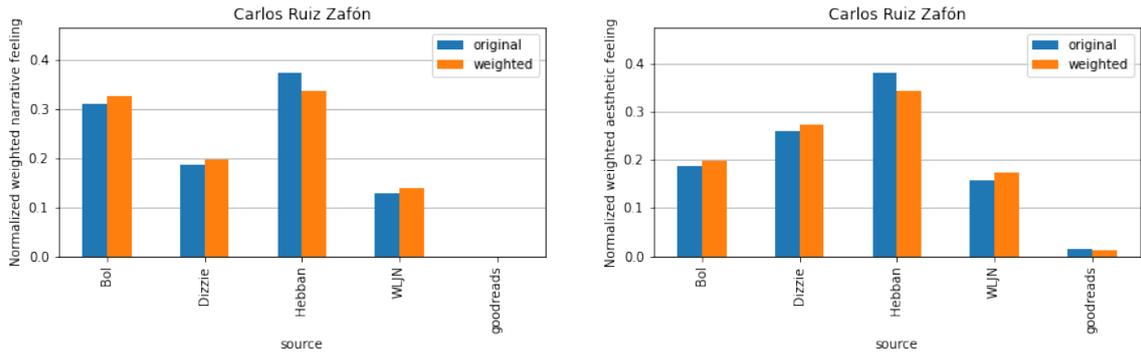


Figure 11: The impact of weighting the number of impact matches by the log-length of reviews for narrative feeling (left) and aesthetic feeling (right) for reviews from different platforms for Carlos Ruiz Zafón's *The shadow of the wind*.

than those of other platforms) and the other platforms becomes smaller. Figure 11 shows this for Carlos Ruiz Zafón's *The shadow of the wind*, but for the other five books, the trend is the same. This suggests that the weighting is effective in reducing differences between review platforms and makes their reviews more comparable. However, these differences across platform remain large, so further investigation is needed into the possibility that reviewers write different reviews for different platforms, either because platforms have different review writing conventions and reviewers modify their reviewing style to each platform, or because the different platforms attract different types of reviewers, who write different kinds of reviews or who experience books differently.

5. Conclusions

This paper provides an in-depth data analysis of the characteristics of online book reviews, to gain insight in how they are related to the reading impact that is expressed in them. Our aim was to find an informed approach to translate impact expressions as identified by the reading impact model [4] on a collection of Dutch online book reviews into an meaningful score so that reading impact can be compared across reviews.

Because collections of online reviews, like other user-generated content on the web, are skewed towards short reviews and popular books, we first analysed how the length of reviews is related to 1) the online platform on which the reviews were published, 2) the number of reviews that a reviewer has written, 3) the popularity of the reviewed books, and 4) book genre. We found that review lengths differ somewhat across platforms, either because of different length restrictions imposed by the platform or different motivations for writing a review on book selling platforms versus social cataloguing platforms, so reviews cannot be straightforwardly compared across platforms without taking these differences into account. There is no substantial difference in review lengths between popular and non-popular books, indicating no underlying length biases when comparing sets of reviews across different books. However, review length is related to the number of reviews that a reviewer has written.

Next, we found that the probability that a review contains an expression of reading impact grows close to log-linearly with the length of reviews, which suggest we should take this re-

relationship into account when comparing aggregate scores per review, book, author or genre. We used this to derive a reasoned method for normalizing the number of impact matches by review length. The impact of weighting is relatively small for books with many reviews, and does not flatten all differences between books, and in some cases makes them more pronounced. However, for different review platforms with different communities of reviewers and different motivations to write reviews, our findings suggest that weighting makes reviews more comparable in terms of scoring impact, although there seem to be more aspects playing a role than length alone. Length normalization reduces only a small part of the differences across platforms. Furthermore, we found that frequent reviewers write reviews that are more consistent in length and in balancing impact expressions related to narrative and aesthetics. This might be a signal that frequent reviewers adopt or create genre conventions [10, 45].

In future work, we want to investigate these frequent reviewers and genre conventions of online book reviews in more detail, as well how they relate to the platforms that the reviews are published on. Another aspect to look at is the types of books that reviewers read and review in terms of popularity and genre, and investigate how reading impact for a genre differs across types of readers.

References

- [1] A. Bachmann-Stein. “Zur Praxis des Bewertens in Laienrezensionen”. In: *Literaturkritik heute. Tendenzen–Traditionen–Vermittlung*. V&R unipress, 2015, pp. 77–91.
- [2] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. ” O’Reilly Media, Inc.”, 2009.
- [3] P. Boot. “A Database of Online Book Response and the Nature of the Literary Thriller”. In: *Digital Humanities*. 2017, p. 4.
- [4] P. Boot and M. Koolen. “Captivating, Splendid or Instructive? Assessing the Impact of Reading in Online Book Reviews”. In: *Scientific Study of Literature* 10 (1 2020), pp. 66–93. DOI: 10.1075/ssol.20003.boos.
- [5] P. Boot, H. Zijlstra, and R. Geenen. “The Dutch Translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary”. In: *Dutch Journal of Applied Linguistics* 6.1 (2017), pp. 65–76.
- [6] J. A. Chevalier and D. Mayzlin. “The Effect of Word of Mouth on Sales: Online book reviews”. In: *Journal of marketing research* 43.3 (2006), pp. 345–354.
- [7] Y. Choi and S. Joo. “Identifying Facets of Reader-Generated Online Reviews of Children’s Books Based on a Textual Analysis Approach”. In: *The Library Quarterly* 90.3 (2020), pp. 349–363.
- [8] d. m. boyd danah m and N. B. Ellison. “Social Network Sites: Definition, History, and Scholarship”. In: *Journal of computer-mediated communication* 13.1 (2007), pp. 210–230.
- [9] S. Dimitrov et al. “Goodreads Versus Amazon: The Effect of Decoupling Book Reviewing And Book Selling”. In: *ICWSM*. 2015, pp. 602–605.
- [10] S. Domsch. “Critical Genres. Generic Changes of Literary Criticism”. In: *Genres in the Internet: issues in the theory of genre* 188 (2009), p. 221.
- [11] B. Driscoll and D. Rehberg Sedo. “Faraway, so Close: Seeing the Intimacy in Goodreads Reviews”. In: *Qualitative Inquiry* 25.3 (2019), pp. 248–259.

- [12] J. Drucker. *Graphesis: Visual Forms of Knowledge Production*. en. metaLABprojects. Cambridge, Massachusetts: Harvard University Press, 2014. ISBN: 978-0-674-72493-8.
- [13] J. Drucker and C. Bishop. “A Conversation on Digital Art History”. In: *Debates in the Digital Humanities 2019*. Ed. by M. K. Gold and L. F. Klein. Minneapolis: University of Minnesota Press, 2019, pp. 321–334. ISBN: 978-1-5179-0692-4. URL: <http://dhdebates.gc.cuny.edu/debates/text/65>.
- [14] E. F. Finn. *The Social Lives of Books: Literary Networks in Contemporary American Fiction (PhD thesis)*. Stanford University, 2011.
- [15] R. J. Gerrig and D. N. Rapp. “Psychological Processes Underlying Literary Impact”. In: *Poetics Today* 25.2 (2004), pp. 265–281.
- [16] P. C. Gutjahr. “No Longer Left Behind: Amazon.com, Reader-Response, and the Changing Fortunes of the Christian Novel in America”. In: *Book History* 5 (2002), pp. 209–236.
- [17] L. Hajibayova. “Investigation of Goodreads’ reviews: Kakutanied, deceived or simply honest?” In: *Journal of Documentation* (2019).
- [18] M. Hundt, N. Nesselhauf, and C. Biewer. “Corpus Linguistics and the Web”. In: *Corpus linguistics and the web*. Brill Rodopi, 2007, pp. 1–5.
- [19] S. Keen. *Empathy and the Novel*. Oxford University Press on Demand, 2007.
- [20] E. M. E. Koopman. “Effects of “Literariness” on Emotions and on Empathy and Reflection after Reading”. In: *Psychology of Aesthetics, Creativity, and the Arts* 10.1 (2016), p. 82.
- [21] E. M. E. Koopman and F. Hakemulder. “Effects of Literature on Empathy and Self-Reflection: A Theoretical-Empirical framework”. In: *Journal of Literary Theory* 9.1 (2015), pp. 79–111.
- [22] M. M. Kuijpers et al. “Exploring Absorbing Reading Experiences”. In: *Scientific Study of Literature* 4.1 (2014).
- [23] P. Lendvai et al. “Detection of Reading Absorption in User-Generated Book Reviews: Resources Creation and Evaluation”. In: *LREC 2020-12th Conference on Language Resources and Evaluation*. 2020, pp. 4835–4841.
- [24] D. S. Miall and D. Kuiken. “A Feeling for Fiction: Becoming What We Behold”. In: *Poetics* 30.4 (2002), pp. 221–241.
- [25] S. Murray. *The Digital Literary Sphere: Reading, Writing, and Selling Books in the Internet Era*. JHU Press, 2018.
- [26] C. Naper. “Experiencing the Social Melodrama in the Twenty-First Century: Approaches of Amateur and Professional Criticism”. In: *Plotting the reading experience: Theory / practice / politics*. Wilfrid Laurier Univ. Press, 2016, pp. 317–331.
- [27] V. Nell. *Lost in a Book: The Psychology of Reading for Pleasure*. Yale University Press, 1988.
- [28] L. Nuttall and C. Harrison. “Wolfing down the Twilight Series: Metaphors for Reading in Online Reviews”. In: *Contemporary Media Stylistics* (2020), p. 35.
- [29] K. Oatley. “A Taxonomy of the Emotions of Literary Response and a Theory of Identification in Fictional Narrative”. In: *Poetics* 23.1-2 (1994), pp. 53–74.

- [30] X. Ochoa and E. Duval. *Quantitative Analysis of User-Generated Content on the Web*. 2008.
- [31] M. Ott, C. Cardie, and J. Hancock. “Estimating the Prevalence of Deception in Online Review Communities”. In: *Proceedings of the 21st international conference on World Wide Web*. 2012, pp. 201–210.
- [32] Z. Papacharissi. “A Networked Self”. In: *A networked self: Identity, community, and culture on social network sites* (2011), pp. 304–318.
- [33] H. Paulussen et al. “Dutch Parallel Corpus: A Balanced Parallel Corpus for Dutch-English and Dutch-French”. In: *Essential Speech and language technology for Dutch*. Springer, Berlin, Heidelberg, 2013, pp. 185–199.
- [34] J. Ratkiewicz et al. “Characterizing and Modeling the Dynamics of Online Popularity”. In: *Physical review letters* 105.15 (2010), p. 158701.
- [35] S. Rebora, P. Lendvai, and M. Kuijpers. “Reader Experience Labeling Automatized: Text Similarity Classification of User-Generated Book Reviews”. In: *Proceedings of the European Association for Digital Humanities Conference 2018 (EADH)*. 2018, p. 5.
- [36] S. Rebora et al. “Digital Humanities and Digital Social Reading”. In: *OSF Preprints* (2019).
- [37] M. Rehfeldt. “Leserrezensionen als Rezeptionsdokumente. Zum Nutzen nicht-professioneller Literaturkritiken für die Literaturwissenschaft”. In: *Die Rezension. Aktuelle Tendenzen der Literaturkritik* (2017).
- [38] C. S. Ross. “Finding without Seeking: the Information Encounter in the Context of Reading for Pleasure”. In: *Information Processing & Management* 35.6 (1999), pp. 783–799.
- [39] G. Sabine and P. Sabine. *Books That Made the Difference: What People Told Us*. ERIC, 1983.
- [40] A. Sairio. “‘No Other Reviews, no Purchase, no Wish List’: Book Reviews and Community Norms on Amazon.com”. In: *Studies in Variation, Contacts and Change in English* 15 (2014).
- [41] D. Smith. “Amazon Reviewers Brought to Book”. In: *The Guardian* February 14 (2004).
- [42] L. F. Spiteri and J. Pecoskie. “Affective Taxonomies of the Reading Experience: Using User-Generated Reviews for Readers’ Advisory”. In: *Proceedings of the Association for Information Science and Technology* 53.1 (2016), pp. 1–9.
- [43] S. Stein. “Laienliteraturkritik—Charakteristika und Funktionen von Laienrezensionen im Literaturbetrieb”. In: *Literaturkritik heute. Tendenzen—Traditionen—Vermittlung*. V&R unipress, 2015, pp. 59–76.
- [44] D. Streitfeld. “The Best Book Reviews Money can Buy”. In: *The New York Times* 25.08 (2012).
- [45] M. Taboada. “Stages in an Online Review Genre”. In: *Text & Talk* 31.2 (2011), pp. 247–269.
- [46] M. Thelwall. “Reader and Author Gender and Genre in Goodreads”. In: *Journal of Librarianship and Information Science* 51.2 (2019), pp. 403–430.

- [47] M. Thelwall and K. Kousha. “Goodreads: A Social Network site for Book Readers”. In: *Journal of the Association for Information Science and Technology* 68.4 (2017), pp. 972–983.
- [48] L. K. Wallace. ““My History, Finally Invented”: Nightwood and Its Publics”. In: *QED: A Journal in GLBTQ Worldmaking* 3.3 (2016), pp. 71–94.