Captivating, Splendid or Instructive?
Assessing the Impact of Reading in Online Book Reviews

Peter Boot[1] and Marijn Koolen[2]
[1] Huygens Institute for the History of the Netherlands,
Royal Netherlands Academy of Arts and Sciences
[2] Humanities Cluster, Royal Netherlands Academy of Arts and Sciences

Author Note
Peter Boot https://orcid.org/0000-0002-7399-3539
Marijn Koolen https://orcid.org/0000-0002-0301-2029
Correspondence concerning this article should be addressed to Peter Boot, Huygens ING, PO Box 10855, 1001 EW Amsterdam, The Netherlands. Email: peter.boot@huygens.knaw.nl

Abstract

What is the impact of reading fiction? We analyse online Dutch book reviews to detect overall affective impact, narrative feelings, response to style and reflection. We create a set of rules that analyze the reviews and detect the impact aspects. We evaluate the detection by asking raters about the presence of these aspects in reviews and comparing these ratings to our detection. Inter rater agreements are weak to moderate; however, there is a significant correlation between the model's predictions for all impact aspects except reflection. The detected impact correlates with book genres in the way one would expect: narrative feelings are highest for thrillers, stylistic response is highest for literary books. We can thus estimate some aspects of the response books evoke in readers. Initial results suggest that the appreciation of style is linked to reflection in the reader. However, the concepts underlying the impact categories need further exploration.

Keywords:
Online book review; reading impact; aesthetic feeling; narrative feeling; reflection; survey

What is the impact that reading fiction has on readers? Does it captivate us, does it cause admiration, does it teach us something? Does it increase empathy (Kidd & Castano, 2013)? Can it cure us (Berthoud & Elderkin, 2013)? We will show some aspects of the impact of fiction in a collection of Dutch book reviews that we downloaded from a number of review sites. We look at four types of impact: overall affective impact, narrative feelings, response to style and reflection. We create a set of rules that, based on the presence of words and phrases in the review texts, determine whether the review shows one or more forms of impact. We evaluate the quality of the predictions by asking contributors of online reviews to rate the presence of these aspects in reviews and comparing these ratings to our predictions. We also show that the predicted aspects of impact are plausible, in that they correlate with book genres the way one would expect: narrative feelings are highest for thrillers and other plot-oriented genres, stylistic response and reflection are highest for literary books.

This means that we have created an instrument that can meaningfully gauge some aspects of the response books evoke in readers. As a first application, we look at the correlation between the different impact aspects, and show that the appreciation of style is linked to reflection in the reader, while narrative engagement is negatively correlated with both stylistic appreciation and reflection.

In this article we discuss in turn the impact of reading and the aspects we investigate, the use of online book reviews and some of its limitations, the model that we designed to capture impact in online book reviews, the survey that we designed to validate the model, the results of the survey and two tentative applications of our model. Finally, we discuss our results, limitations and prospects of what we have done.

Reading impact

There are many ways in which reading stories can impact readers. The most important effects are often thought to be emotional. Miall and Kuiken (2002) find four types of feeling as a result of reading: enjoyment in reading, empathy or sympathy, aesthetic response, and feelings that modify the self. They also distinguish between fresh and remembered emotions. Oatley (1994) contributes the distinction between ‘external’ emotions (the reader confronting the text) and ‘internal’ emotions (the reader entering the world of the text). Other possible effects are personal transformation (mentioned by Miall and Kuiken as well as Sabine and Sabine 1983, Ross 1999), (self-)reflection (Koopman 2015), increased empathy (discussed in Keen, 2007) and beliefs about the real world (Gerrig and Rapp 2004).

In this paper we do not intend to look at all aspects of reading impact. Our intention is to explore how the availability of online reviews might contribute to its study. We take a computational approach: we want to develop a tool that will be able to decide, for a given review, what forms of reading impact are present in the review. Basing ourselves on Koopman and Hakemulder (2015) we choose to focus for now on four aspects of impact: general emotional impact, narrative feeling, aesthetic feeling and reflection. We know these aspects are well represented in the reviews and probably easier to detect than other potential responses, no

less interesting (such as personal transformation), that we may consider at a later stage. We conceptualize narrative and aesthetic feelings as subcategories of general emotional impact. Narrative feelings and response to style might be related to increased empathy, according to Koopman (2016), which might, together with reflection, be related to the prosocial effects of reading. Apart from their intrinsic interest, these aspects of impact might therefore also be of social importance. We define them as follows:

- emotional impact: any (fleeting or permanent) emotional response during or after reading, be it targeted at the book as a whole, its author, its characters, its style, aspects of the world that it describes, or any other aspects. These emotions may be positive or negative, strong or weak. We exclude, however, emotions such as boredom, where the book simply has failed to work;
- narrative feeling: any feeling with respect to the narrative world and the characters; this includes both being drawn into the narrative world (absorption or transportation) and character-directed emotions such as sympathy and identification (Koopman and Hakemulder 2015, p. 90);
- aesthetic feeling: any feeling targeted at the aesthetic features of the text, such as admiration, appreciation, surprise and defamiliarization (Koopman and Hakemulder 2015, p. 94);
- reflection: thoughts, insights and musings on oneself, others, society or the book. Unlike Koopman and Hakemulder, we do not require reflection to be about the self.

We note that impact, as we define it, is not something that necessarily lasts much longer than the reading experience. Neither do we require it to transcend reading, in the sense that it should affect domains of life other than reading.

Online book discussion

We study reading impact as it appears in online book reviews. Our study is not experimental but is meant to complement existing research with data obtained 'in the field', based on large numbers of readers reading the books they chose themselves in natural circumstances. We first review some of the existing literature on online book discussion, then we discuss some of its limitations

Related research

Online book discussion is a multifaceted phenomenon. It can take the form of book blogs (Steiner 2010), of book discussion on Twitter (Gruzd, Rehberg Sedo, 2012), in Instagram (Jaakkola 2019), of comments on classical works on Wattpad (Pianzola et al. 2020), of reviews on Amazon and other booksellers (Mehling, Kellermann, Kellermann, & Rehfeldt, 2018), of reviews and discussion on review-and-networking sites such as Goodreads (Thellwall & Kousha, 2016) as well as many other text genres and platforms (Boot, 2011). Research into online book discussion was initially often motivated by marketing purposes (Chevalier & Mayzlin, 2006). Other researchers explored the consequences of online book discussion for

book-related professions, such as libraries (Naik, 2012) and the book trade (Martens, 2016; Murray 2016). Murray also looked at the implications of online book discussion for the discipline of book history (2018). Much research has been done on the community aspects of book discussion sites (Worrall 2019) and how networked reading influences the reading experience (Albrechtslund, 2019). Especially in Germany, researchers have looked at the characteristics of online reviews vis-à-vis professional reviews (e.g. Rehfeldt, 2017a). Some have deplored the diminishing respect for the professional critic that online book discussion demonstrates (McDonald, 2007).

Most of these researchers have looked at the properties of the platforms and the ways that the networked setting influences communication and reviewing on the platform. The number of researchers that have used these platforms to see what they can teach us about reading and literary reception more generally (Rehfeldt, 2017b) is much more limited. Gutjahr was probably the first one to use Amazon reviews for studying literary reception, in his investigation into American protestants' uptake of the *Left Behind* novel series (2002). Finn uses both reviews and Amazon recommendation ('also bought') data in his study of (the reception of) US contemporary authors (2011). In recent times, the interest in online book discussion for reception studies is increasing. In some cases researchers collect a limited number of reviews for studying the response to a single book, author or genre: Wallace looks at a number of Goodreads reviews of Barnes' *Nightwood* to look at continuities between early and modern reception (2016), Naper uses 1500 Norwegian user reviews in her study of what she calls the 'social melodrama' genre (2016).

While there is nothing wrong with studying individual reader reviews and building an argument based on them, the available number of reviews suggests completely different methodologies for studying response, based on computational analysis. Ridenour and Jeong, for example, coming from a library background, create clusters of co-read books on Goodreads (books read by the same user) (2016). Hajibayova looks at some linguistic characteristics and their psychological implications in 475,000 Goodreads reviews (2019). Thelwall looks at effects of reader and author gender in fifty different genres based on some 200,000 Goodreads reviews (2019). What these publications have in common is that they have made visible aspects of reading behaviour that have been largely hidden up to now. Driscoll and Rehberg Sedo (2019) combine (manual) content analysis with (machine) sentiment analysis in their investigation of emotion and sociality in Goodreads reviews. Their study is close to ours in the sense that they are interested in aspects of the reading experience (they code for temporal, intellectual, emotional and physical aspects) that are related to the aspects that we look at. Other work that uses online reviews for computational analysis of reading is reported in (Rebora, Lendvai & Kuijpers, 2018), who explore the feasibility of using online reviews to validate the aspects of the Storyworld Absorption Scale (Kuijpers, Hakemulder, Tan & Doicaru, 2014).

Limitations

From the outset, however, we have to make a number of caveats. We are researching the impact of reading. The source for our data is reader reviews. This presumes a number of things: (1) that the contributors of online reviews are somewhat representative of the general or ordinary reader; (2) that the reviews they contribute are sincere (not fake); and (3) that

performative considerations, social desirability and genre expectations are not overriding factors in the creation of the reviews. We address each of these issues briefly.

With respect to representativeness (1), there are no doubt important differences between 'regular' readers and readers that contribute book reviews. The reviewers are probably heavier and more experienced readers. Van Putten-Brons & Boot (2017) found that most people who wrote about Dutch literature in English were mostly well-educated. Similar results were found by Toth and Audunson (2012) in their study of Norwegian and Hungarian sites. There is no way of knowing what the reviews of the non-reviewing readers would look like, and we should therefore be cautious in drawing conclusions about book impact on all readers. That said, we will see below that the impact detector we describe does succeed in detecting interpretable and consistent differences between the reviews of different genres. Regardless of the extent of the reading population that these measurements are based on, they do show differences in book impact that need explanation.

With respect to the reviewers' sincerity (2), it is well known that some authors try to game the reviewing system by positively reviewing their own books and negatively reviewing those of their competitors (Smith, 2004). Authors can even order collections of reviews (Streitfeld, 2012). While there are some tell-tale characteristics of fake reviews created by naive fraudsters (Sairio, 2014), it is very hard to assess the prevalence of fake reviews. One study suggests percentages of below six percent in six different review communities (Ott, Cardie, & Hancock, 2012). Based on this study we will assume that fake reviews are not so frequent as to make the average review untrustworthy.

Finally, sociologists and new media researchers have shown (3) that social media behaviour is not just a straightforward expression of pre-existing opinions. Building on the ideas of Goffman (*The Presentation of Self in Everyday Life*, 1956) social media behaviour has been studied as a form of identity performance (boyd and Ellison, 2008; Papacharassi, 2011: 304). More specifically reviews (on Tripadvisor) have been shown to 'construct identities as particular types of individuals in this online context' (Vázquez, 2014). There is, as far as we know, no research confirming this effect specifically for online book discussion. However, it is certainly plausible that book reviewers, while sharing their view of a book, are at the same time working to project an image of themselves as intelligent, sweet-tempered, merciless, or any other property they would like to be seen to possess. Beyond these social considerations, reviews are also a genre, whose conventions are changing with the move to the Internet (Domsch, 2009; Stein, 2015). Reviews are written with genre conventions in mind (Bachmann-Stein, 2015; Taboada, 2011). Again, it is impossible to know what would be the difference between the reviews as we have them and the way reviews would have turned out without the identity work and the genre expectations. We assume the effect is not inhabilitating for what we are trying to do, but it is clear that there is much more to be studied in the reviews than the aspects that we are interested in.

The impact predictor: development

In this paper we look at how we can determine impact from the online book reviews that readers write on sites such as Amazon or Goodreads, or in the Netherlands bol.com or hebban.nl. In

many online book reviews, readers clearly express impact. They write things like 'It is easy *to enter into* Daniel's doubts and confusion', 'It *was written beautifully and compelling* without becoming melodramatic' or 'Takano knows how to render a story that makes a lasting impression, that is *instructive and gives food for thought* (...)'.[1]

We created an impact predictor for Dutch book reviews that consists of three related resources: (i) a list of potential impact terms, taken from a number of existing resources as well as based on what we found in the reviews, (ii) a list of book aspect terms that groups e.g. words related to style or words related to plot, and (iii) a set of rules that associate impact terms with specific aspects of impact, possibly referring to word groups defined in the list of book aspect terms. The rules are evaluated based on individual sentences of the review; if the conditions are met, the mentioned impact aspect is assigned to the review sentence. To be clear: we do not expect our rules to take into account the full text of the review. We are only interested in the phrases that signal impact of reading on the reviewer. We describe the data we use and each of these resources.

Data

The book reviews that we used in developing and evaluating our predictor are those from the Online Dutch Book Response (ODBR) corpus (Boot 2017). The corpus contains reviews from a number of Dutch book discussion sites (hebban.nl as well as a number of sites that have disappeared) as well as from bol.com, one of the biggest Dutch online booksellers. We used two sets of reviews: the first set for developing the rules and exploring the usefulness of the predictor; the second set for evaluating the predictor. See table 1 for a summary. The texts have been Part-Of-Speech (POS) tagged and lemmatized using Alpino (Van Noord, 2006).

Table 1
Data used in developing, evaluating and applying the predictor

Label	Number of reviews	Sources	Description	Used for
Set 1	382208	Hebban.nl, bol.com and other sites	Large collection of reviews; includes Hebban reviews with date <= 2016-06-12	Used to look for impact terms and suitable rules. Used also for looking at the predictor scores for individual books and for computing correlations between the different aspects
Set 2	2743	Hebban.nl	Random selection from reviews downloaded from Hebban with date > 2016-06-12	Comparing survey results with model predictions.

[1] https://www.hebban.nl/recensies/mads-bruynesteyn-over-de-boerderij, https://www.hebban.nl/recensies/manja-4530-over-geluksvogel, https://www.hebban.nl/recensie/edwin-lommers-over-executie, all consulted November 10, 2019.

Impact terms

The impact terms are words and phrases that might indicate the presence of impact in a text. The list now contains ca. 1100 entries. It is divided into thirteen categories (table 2). The adjective_terms and noun_terms can be either book or reader related ('scary' vs. 'afraid'). These categories, as well as the verb_terms are interpreted as lemmas. A verb_term such as 'enjoy' will be evaluated by testing whether the word in the text has lemma 'enjoy' and POS-tag 'verb'. The phrases, in contrast, are evaluated on the tokenized (split into words) representation of the review's text. They consist of a sequence of words possibly with sets of alternatives for certain positions in the phrase (for example: 'at|on' the edge of (my|your|one's) (chair|seat)'). In the discontinuous phrase, unlike in the continuous ones, extra tokens may be present in between the tokens that are part of the pattern. Partic_terms are used for participles used as adjectives, as parsers may consider these as either verbs or adjectives.

Table 2.
Categories in impact list.

adjective_term_book_related, adjective_term_reader_related, noun_continuous_phrase, noun_discontinuous_phrase, noun_term_book_related, noun_term_reader_related, other, other_continuous_phrase, other_discontinuous_phrase, partic_term, verb_continuous_phrase, verb_discontinuous_phrase, verb_term

The terms in the list were derived from a number of sources (Boot 2012, Saricks 2005, Knoop et al. 2016, Schindler et al. 2017, Hosoya et al. 2017, Kuijpers 2014, Dal cin, Zanna & Fong, 2004 (as reported in Kuijpers 2014), Knobloch-Westerwick & Keplinger, 2006 (as reported in Kuijpers 2014), Busselle & Bilandzic, 2009 (as reported in Kuijpers 2014), Appel et al. 2002, Spiteri and Pecoskie 2016). Where available we used the translation into Dutch as provided in Kuijpers 2014, if not, we made our own translation. A first check of many terms was done using searches in Set 1 of the ODBR data. These searches also resulted in many potential new impact terms.

The entries in the list can indicate any sort of impact, such as strong emotional impact (*hartverscheurend*, 'heartrending'), absence of surprise (*voorspelbaar*, 'predictable'), tones (*enthousiast*, 'enthusiastic'), evaluative terms (*mooi*, 'beautiful'). The impact term itself does not indicate which type of impact a term indicates, as that may depend on the context. This explains the need for the rules and the aspect terms described below: if an impact term is encountered in a review, the rules and aspect terms will determine what type of impact applies. Impact terms are also very different with respect to their precision: *meeslepend* ('compelling') will probably always indicate impact, *slecht* ('bad') will only do so in a limited number of contexts. The phrases in the list usually pre-select the relevant contexts: 'see' in itself would not indicate impact, but the combination *voor (me|mij|je) (zien|ziet|zag)* ('(see|saw) in front of (me|you)') probably does.

Book aspect terms

The list of book aspect terms is used to determine what aspect of a book or a reading experience a certain impact term probably refers to. It contains words (lemmas or patterns) associated with the following book aspects: general, author, reader, plot, character, setting, style and subject. For instance, the (Dutch equivalents of the) words 'I', 'me', 'you', 'reader' are associated with the aspect *reader*: if the word *genieten* ('enjoy') occurs in the same sentence as one of the words from the reader category, it probably refers to the reviewer enjoying something (rather than a character). Words associated with the 'general' aspect include the patterns **boek* ('book'), **roman* ('novel') and *debut* ('debut'). Some of the aspects are loosely based on the appeal categories mentioned by Saricks (2005).

Term - Impact Rule set

The rule set associates impact terms with impact aspects, possibly under a certain condition. Table 3 is a small extract of the rule set.

Table 3.
Extract of rule set

Impact_group	Impact_term	Code as	Condition	neg filter
partic_term	spannend	N		
adjective_term_book_related	prachtig	S	@style	
verb_term	verplaatsen	N	%in	
verb_term	beschrijven	S	%beschrijft	y

In the first line, the word *spannend* ('suspenseful') is unconditionally associated with N (narrative feeling). In the second line, the word *prachtig* ('beautiful') is associated with S (stylistic feeling). The 'condition' column adds a requirement that the word should appear in the same sentence as one of the 'style' aspect terms. To give an example: the sentence 'It was a beautiful and happy ending, which I love', wouldn't qualify, but 'Her language in these paragraphs was beautiful', would, because the word 'language' is one of the style aspect terms. The condition column uses a number of special characters in order to indicate different sorts of conditions. The third line shows an example that associates the verb *verplaatsen* ('project oneself', 'enter into') with narrative feeling if the word *in* ('in') appears in the same sentence (otherwise *verplaatsen* might have a completely different meaning). In the last line, the verb *beschrijven* ('describe') is associated with style, here with a negative condition (the 'neg filter' column has value 'y'): the association only holds when the word *beschrijft* (third person singular of *beschrijven*) does *not* appear in the sentence (as that form of the verb is usually used in a factual, non-evaluative context).

These rules were written manually, based on inspection of the contexts of the impact terms in the ODBR corpus. For the 1100 terms in the impact list, we computed their frequencies in the corpus, and created rules for the 250 most frequent terms. Creating rules for the less frequently occurring terms would have been prohibitively time-consuming.[2] The rule set is thus far from exhaustive: many phrases that would indicate a form of impact are not represented in

[2] We estimate that the creation of the rules took ca. 80 hours.

the rule list. There is certainly room for improvement in that respect. However, as we started with the most frequent terms, the quantitative impact of adding more rules might be limited. The 1100 impact terms together occur 1.8 million times in the corpus, of which 1.6 million (87%) are of the top 250 terms. We assume diminishing returns with additional effort.

This process resulted in 275 rules for 184 impact terms, as some of the terms were too vague or had different meanings that were hard to distinguish by context. The rules are not evenly distributed over the four categories, with 118 rules (43%) for narrative feeling, 60 (or 22%) for aesthetic feeling, 55 (20%) for emotional impact and 42 (15%) for reflection. The rules are also not evenly distributed when we take the frequency of the terms into account. For the top 50 most frequent impact terms, there are 51 rules, with 20 rules (39%) for aesthetic feeling, 14 rules (27%) for narrative feeling and emotional impact each, and only 3 rules (6%) for reflection. The reflection category is the least well represented in the rules, with only a small number of impact terms and most of them in the lower end of the frequency distribution. This may be because reflection might occur less frequently than the other impact forms, or alternatively because reflective thoughts can be related to many different aspects of the reviewer (their thoughts, their surroundings, their past), of the book or the outside world, and can be worded in many different ways.[3] Given these statistics, we expect our model to predict book reviews to frequently express aesthetic impact and narrative feeling, but that it will struggle to capture expressions of reflection.

Issues

Rule creation was not an easy process. Phrasing suitably generic rules that associate impacts with correct contexts is a challenge, for a number of reasons. The most immediate problem is that words occur in multiple senses, in various grammatical constructions, and are used to refer to various forms of impact. For many words, the contexts were so variable that it wasn't possible to write rules for them. Examples are *kwaad* ('angry') or *puur* ('pure'). *Puur* can certainly be used to indicate uncorrupted characters or narrative, but there is no consistent context that can distinguish these cases from the word's other uses ('unadulterated', 'just', 'alone').

A more fundamental problem is that it is not always clear what form or intensity of impact a certain group of words represents. Humour is a good example: some humour is clearly stylistic, but is humour always a question of style? Can't events be funny too? But if they are, does that make the selection of events also a matter of style? Without doubt, to some extent it is. Another example concerns reflection: the word 'surprise' seems to imply a previous expectation that is being challenged by events, and therefore could constitute an indication for reflection. But is this indication strong enough?

Finally, in many cases, the context of a single sentence may be insufficient to decide whether a term indicates impact. In some cases, we decided to go by plausibility, and for instance created a rule that associates *huilen* ('to weep') with narrative feeling, even though weeping could also be a response to style.

[3] Note that many rules require the impact term to co-occur with a book aspect term, which also have different frequencies, so a rule for a highly frequent term might still be rarely triggered because it requires some low frequency aspect terms.

Because of these issues it is clear that the predictor we developed will not be a hundred percent correct in all cases. What we aim for, however, is a tool that will be right most of the time, that will be able to track patterns in large collections of reviews, even when making mistakes at the level of the individual review. In the next section we will discuss how we validated the predictor.

The impact predictor: evaluation

Survey

The impact predictor associates review texts with impact categories. We should evaluate these predicted categories before using them in further study. That means: we need to test whether, for the audience for which these reviews were written, our predictor correctly classifies the reviews: if we predict a review expresses reflection in the reviewer, does the target audience agree with that prediction?

In order to perform this evaluation, we created a web-based survey, where we asked users to rate sentences in reviews in terms of indications for emotional impact, narrative feeling, aesthetic feeling and reflection. For a sample question, see figure 1. We selected a random sample of 2743 sentences from Hebban reviews to be rated in this way. In order to ensure that our results would not be artificially high because of overfitting to existing data, we only used reviews that had not been consulted in creating the predictor (set 2 from table 1).

The survey was tested by asking a number of colleagues to fill it in and asking them for comments. We received some hard to reconcile remarks: on the one hand, testers felt daunted by the amount of instruction and explanation. On the other hand, testers asked for more explanation about aspects that they didn't quite understand. We tried to satisfy both groups by providing a single page with the information that we felt the absolute minimum, and optional extra pages which we encouraged people to read.

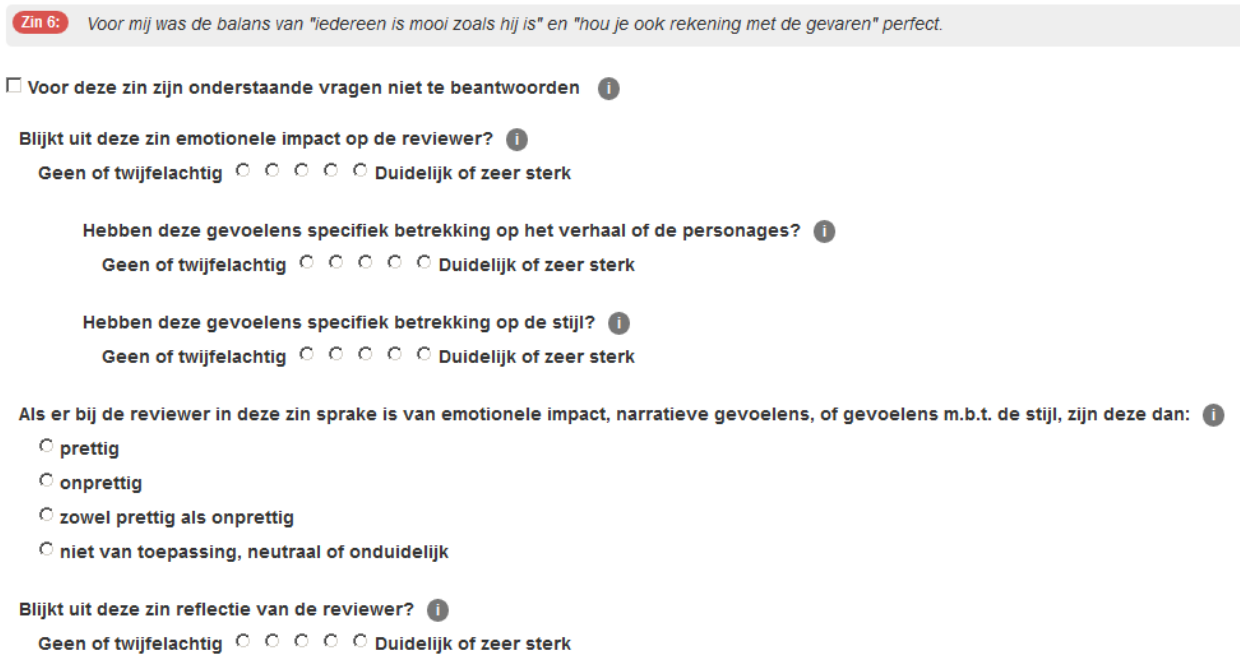

Figure 1. Sample survey question. The question shows a random sentence from a review ('For me, the balance between "everyone is beautiful as he is" and "will you take the dangers into account" was perfect'), followed by items that ask the user to rate whether the test sentence shows emotional impact, narrative feelings and stylistic feelings in the reviewer. The Likert scale items range from 'None or doubtful' to 'Clearly or very strongly'). The next item asks the user to indicate whether these feelings are pleasant, unpleasant, both or none (not used in this article). The last item asks whether the test sentence shows reflection on the part of the reviewer.

We partnered with book review site hebban.nl, the largest Dutch book discussion community,[4] in order to find users willing to take the survey. The survey was set up so that each sentence will be rated by at least three users. Users were presented with a set of 10 sentences to assess. As soon as a user rated a sentence on all aspects, the ratings were stored in the database. Upon rating all 10 sentences, they could quit the survey or opt to rate additional sentences. Users identify themselves to the survey using a server assigned ID, but no personally identifiable data.[5]

Survey results

The survey went live on February 9, 2019. It was announced on Hebban and in the site's daily mailing. On the announcement page[6] some of the people who had done or tried the survey responded. Initially the comments were mostly negative: 'What a strange survey, I quit after sentence three', 'The sentences were unclear and so was the explanation'. We tried to explain once again what the idea was. Later some people also responded positively: 'Pretty fun to do and to think about these sentences', 'Twenty sentences rated, nice to do. Maybe try a few more

[4] At the time of writing it claims 170.000 registered members. See https://www.hebban.nl/promotie
[5] We need the ID in order to make sure that we do not ask a user to rate the same sentence twice.
[6] https://www.hebban.nl/artikelen/wat-voor-effect-hebben-boeken-op-lezers-een-onderzoek-van-het-huygens-instituut

tomorrow'. One person complained about the quality of the language in the sentences, another one about a supposed requirement to register to participate in the survey.

In all, 348 sentences were rated by at least three raters from a total of 109 different raters. 43 participants did not complete the first page in the survey, one participant got to 60 sentences. For the distribution, see figure 2.

Raters could indicate if they could not judge a sentence at all. This happened mostly for very short sentences that are uninterpretable without the rest of the review as context. Of the 348 sentences, 15 were marked as such by at least two of the raters, so no agreement can be computed. These are left out of the rest of the analysis.

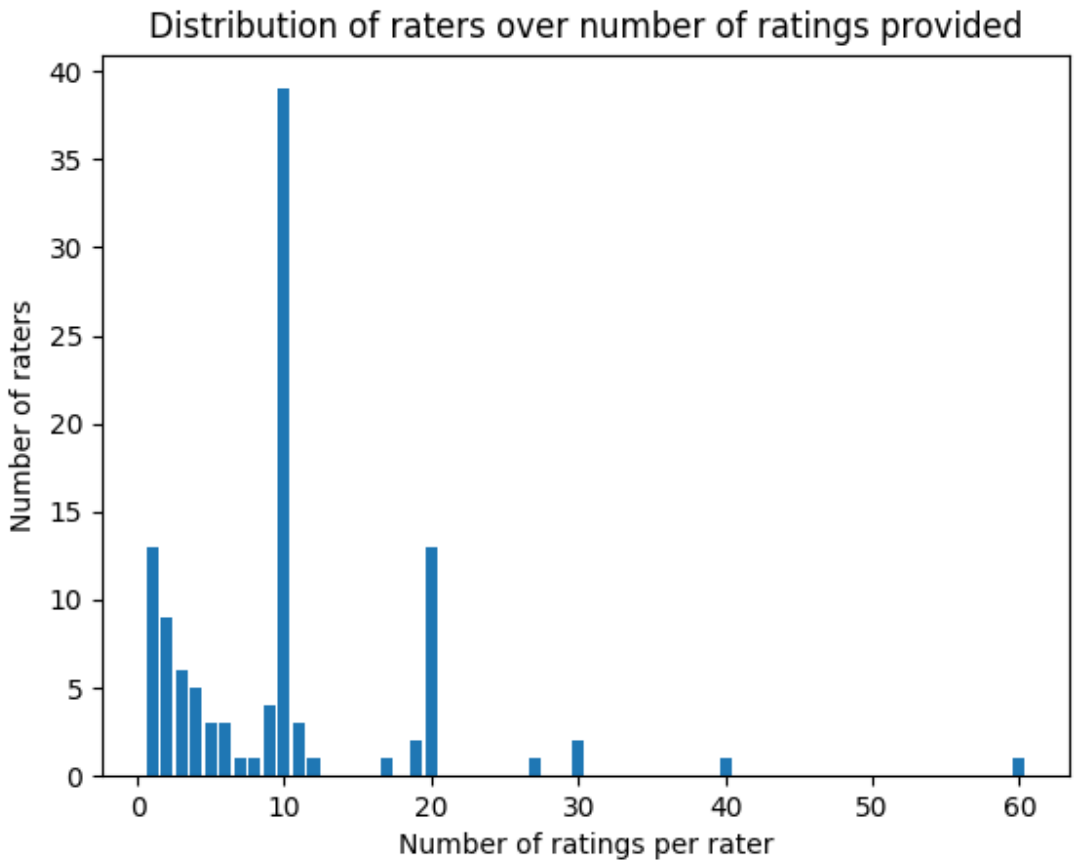

Figure 2. Distribution of raters by number of ratings. Most raters rated ten sentences (one page).

Interrater Agreement

To see if raters interpret the rating task in the same way, we computed Interrater Agreement (IRA).[7] Following the review of IRA measures by O'Neill (2017), we use the IRA statistic r^*_{wg} (Lindell and Brandt 1997). The r^*_{wg} takes values in the [-1,1] interval and negative agreement scores are interpreted as a form of disagreement between subgroups of raters. Although there are some concerns that Likert scale ratings cannot be interpreted as interval data (Jamieson 2004), others have pointed out that in many cases the errors produced by treating them as such are minimal (Norman 2010).

The statistic is computed as:

$$r_{wg} = 1 - \frac{S_x^2}{\sigma^2}$$

where S_X^2 is the variance of the ratings for a sentence and σ^2 is the expected variance based on a chosen theoretical null-distribution that represents a total lack of agreement. LeBreton and Senter (2008) argue that the choice of null-distribution should be guided by the specifics of the experiment and the biases in responses. The most used null-distribution is the uniform null, which assumes that all ratings are equally likely to be chosen. As shown in Figure 3, the 3896 ratings in our show a tendency towards the extremes. On this basis we decided to use an 'inverse triangular' distribution with an expected variance of 2.55 as our theoretical null in calculating agreement, such that agreement scores fall in the range [-0.57, +1].

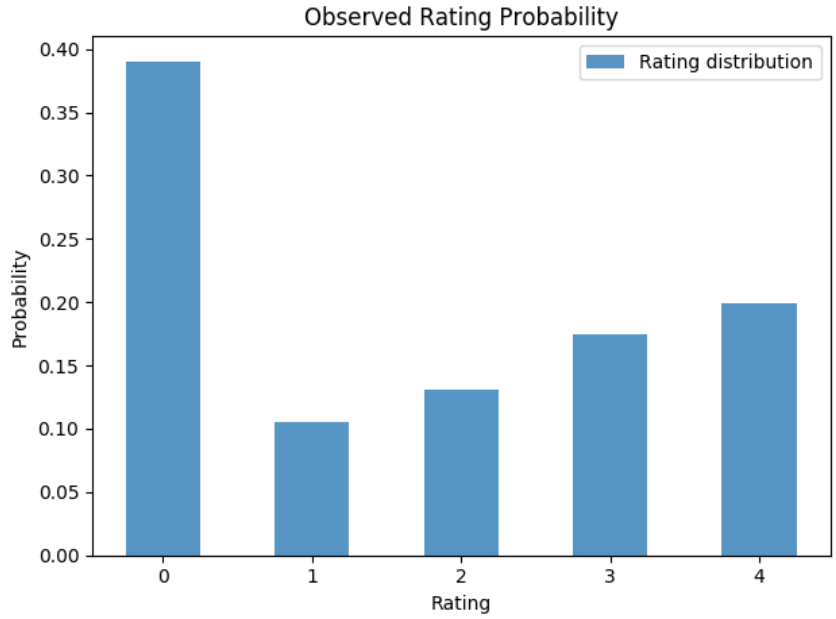

Figure 3. Rating probability distribution over all 3896 ratings across all impact categories.

[7] All data and computations for the sections Interrater Agreement and Rater-model agreement are available in this Github repository: https://github.com/marijnkoolen/reading-impact-agreement-analysis

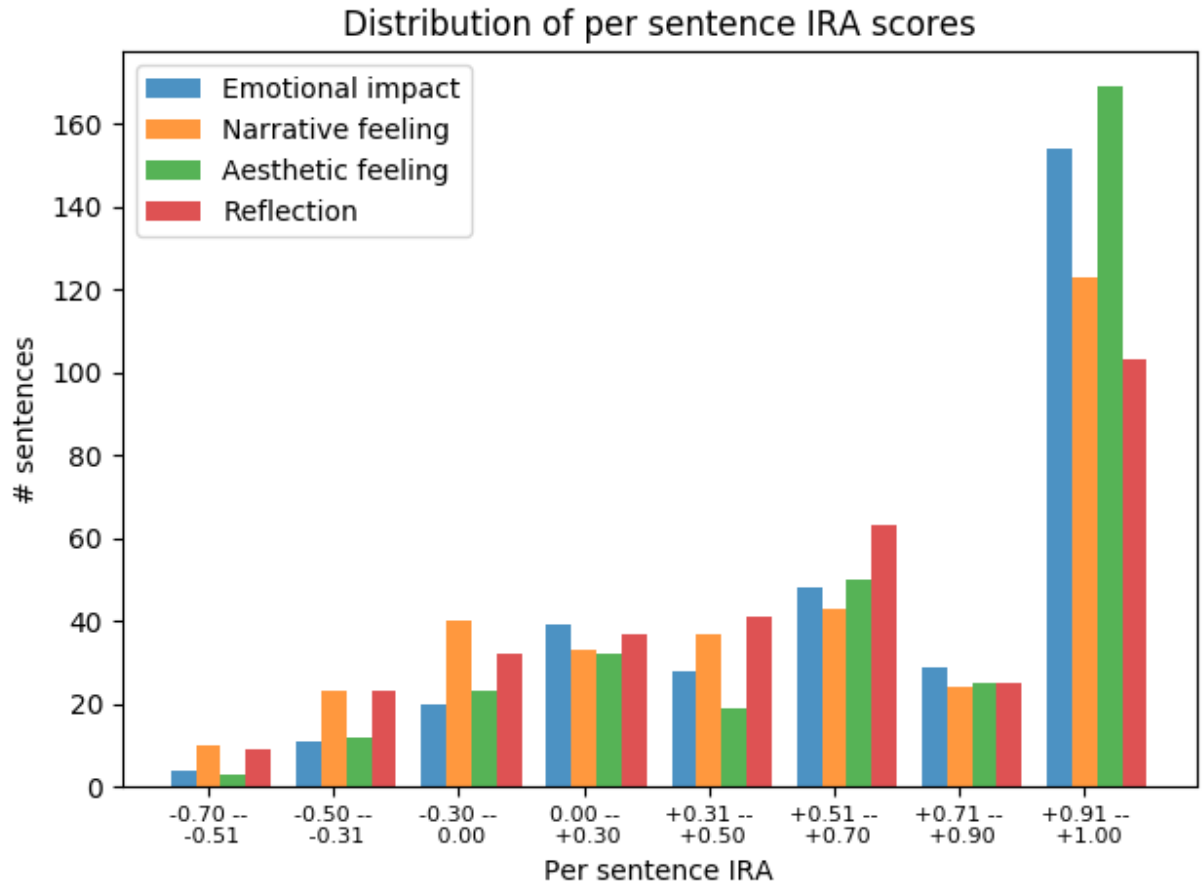

Figure 4. Distribution of IRA scores per sentence, for the four impact categories.

The distribution of IRA scores per sentence and impact category is shown in Figure 4. The average IRA is moderate for emotional impact (0.62) and for aesthetic feeling (0.65), and weak for narrative feeling (0.49) and reflection (0.48).[8] We see two main explanations for this lower agreement for the latter two. One explanation is that raters did not have a concrete enough idea of what each category meant: these terms are essentially scholarly categories that regular readers are not necessarily familiar with. It is also possible that the individual sentences did not give enough context to interpret the review author's thinking, forcing the raters to fill in the gaps with their own interpretations. This last explanation is somewhat supported by the fact that raters indicated for some sentences that they could not judge them on these categories.

Rater-model agreement

[8] We use the guidelines given by a.o. LeBreton and Senter (2008), in which 0-0.30 signals no agreement, 0.31-0.50 weak agreement, 0.51-0.70 moderate agreement, 0.71-0.90 strong agreement and 0.91-1.00 very strong agreement.

After having looked at the agreement between raters, we can now proceed to check the agreement between our impact prediction model and the human raters. In this comparison, we look only at the sentences for which the human raters show moderate agreement or better, i.e. where IRA >= 0.5. What we want to know is: do human raters give a higher rating to sentences that our model predicts as expressing impact than to sentences for which the model thinks there is no impact? The results are shown in table 4: the number of sentences with IRA above the threshold of 0.5, the sentences with and those without predicted impact and the p values of the Mann-Whitney U significance test that compares the model's predictions and the ratings. For the emotional, narrative and aesthetic categories, our model is effective at predicting the human ratings ($p << .001$). As we discussed above, our model has fewer rules for reflection than for the other categories, and those reflection rules tend to target less frequent words. Not surprisingly, the table shows that there are few sentences for which the model predicts an expression of reflection. As a consequence, for the reflection category the validation failed ($p = 0.14$).

Table 4.
Statistics on the number of sentences within each impact category with IRA >= 0.5 (column 2), for which our model predicts no impact (Model = 0, column 3), or does predict impact (Model >= 1, column 4) and the Mann-Whitney-U p-value (column 5).

Category	# sentences IRA>0.5	Model = 0	Model >= 1	Mann-Whitney-U (p-value)
Emotional impact	231	167	64	$9.1 * 10^{-8}$
Narrative feeling	190	165	25	$2.0 * 10^{-5}$
Aesthetic feeling	244	225	19	$3.4 * 10^{-11}$
Reflection	191	185	6	0.14

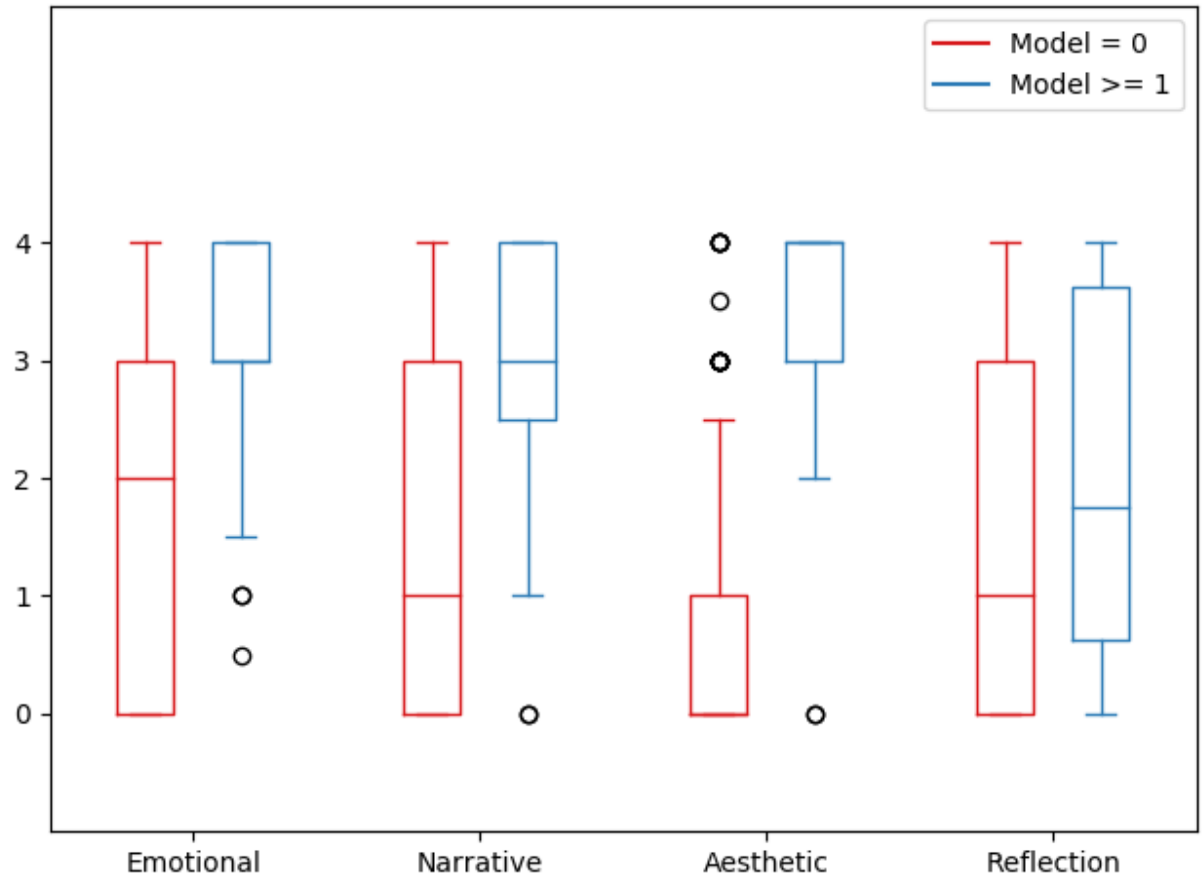

Figure 5. Box-and-whiskers plots for the distribution of median human ratings for each type of impact, i.e emotional impact, narrative feeling, aesthetic feeling and reflection. This is based on sentences with a minimum interrater agreement of r^*_{wg}>=0.5.

Figure 5 gives a visual illustration of the effectiveness of the predictor model, showing the distribution of the median human rating per sentence (so the median of the three ratings per sentence), for each type of impact, split over sentences where our impact predictor identifies impact (Model >= 1) and sentences where it does not (Model = 0). Each box represents the 2nd and 3rd quartile of the ratings, with a line inside the box representing the median rating. The whiskers represent minimum and maximum non-outlier ratings, and the circles represent the outliers. For some boxes, the median is the same as the inter-quartile boundary. For instance, for the emotional impact and aesthetic feeling scales and Model >= 1, the median is at the upper bound (the maximum rating of 4), so the upper two quartiles are also at 4. In other words, if our model detects impact for these categories, it tends to correspond with a human rating of 4, clearly indicating the connection between the model's predictions and the human ratings. Similarly, for aesthetic feeling and Model = 0, the median and therefore the lower two quartiles are at the lower bound (minimum rating), meaning that for the majority of sentences for which our model does not detect impact, human raters give a rating of 0.

Aesthetic feeling is the only impact type for which the distributions in Figure 5 are clearly separate, implying that the model and the raters mostly agree. For the other impact types, the

distributions overlap. For some sentences the model predicts impact that raters disagree with; there are, however, more sentences that are rated as clearly expressing impact by human raters, but that are missed by our model. Perhaps these other types of impact are expressed using a more varied vocabulary and more rules are needed for our model to detect them.

Looking at the cases where there are large differences between the subjects' ratings and our model's predictions, we find different explanations. In a limited number of cases, we indeed encounter words or phrases that our model might have contained, such as 'sucked into the story' or 'worst nightmare' as indicators for narrative involvement. More often, a formulation shows for example reflection, without the presence of a corresponding impact term: when a book's story makes a reviewer ask 'After a war, when are you right, when wrong?', this exemplifies reflection but the sentence contains no explicit reflective phrases. A more frequent reason for disagreement, however, is interpretation: when the reviewer wrote 'I'll certainly read another book from this writer and series for comparison' it may be a safe prediction that the reviewer was absorbed in the narration and that is probably how the raters reasoned, but the text doesn't say so. In other cases, our model does the interpretation, rather than the raters: in the case of 'the book knew how to capture me from start to finish', the model, based on the plot term 'finish', assumes it is narrative interest that is responsible for 'capturing' the reviewer, but strictly speaking we don't know. There are also many boundary cases, where a term in the model indicates a small level of the relevant quantity: the model assumes reviews with the word 'mysterious' show narrative feeling; the subjects didn't always feel that way. Similarly, for the model, 'interesting' is assumed to show reflection. It is clear that, if it does so, the reflection may be only superficial. Finally, there are many cases where apparently the raters have not understood the concepts that we meant to use, as when raters miss the narrative absorption in 'the cold cuts through you like a knife' or 'you could grab the fish in the rivers with your hands', or, the other way around, they see stylistic feeling in 'When I ran into this book she wrote, I had to take it home'.

Rule coverage

We also looked at the fraction of reviews that match our impact rules, to get insight (the distribution of) the number of matching rules per review. For this, we counted the number of matching rules of each impact category per review, using a sample of 50,665 reviews for the books for which we have at least 100 reviews (described in more detail in the next section). In Figure 6 the distribution of number of matching rules per review is shown for all four impact categories individually and combined ("All"). For Aesthetic feeling and especially for Reflection, the majority of reviews match no rules, 20% resp. 15% match one or more. For Emotional impact and narrative feeling the distribution is much less skewed, with just over 40% of reviews match no rules, over 30% match one rule in each category and around 25% of reviews matching two or more rules. Although our model certainly does not detect all expressions of impact, this distribution shows that at least a quarter of reviews contains multiple expressions of narrative and emotional impact. The distribution of all impact categories combined shows that there are very few reviews (13%) for which our model finds no matching impact rules. Another 22% match one rule and over 65% of reviews match at least 2 impact rules, showing that our model is capable of finding expressions of impact in the majority of reviews.

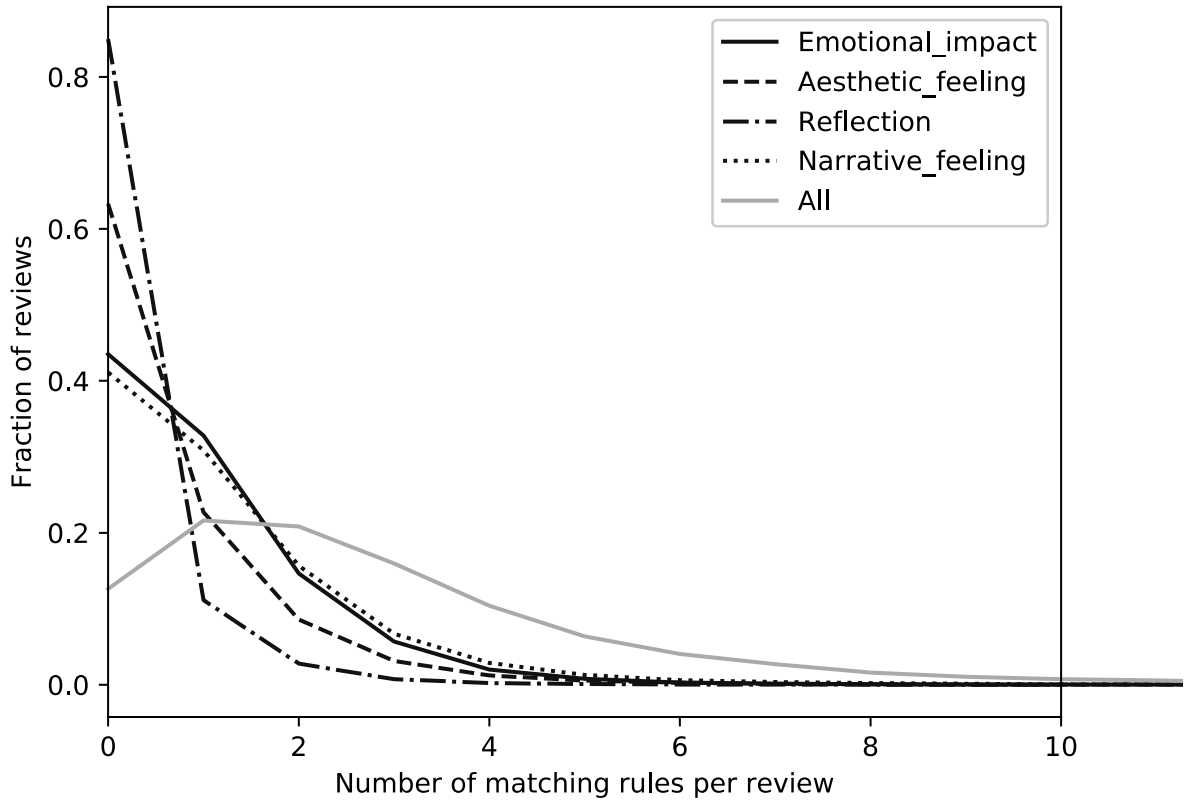

Figure 6. The distribution of the number of matching rules per review for the four impact categories individually and combined. The Y-axis shows the fraction of reviews from a sample of 50,665 reviews.

First results

We did an analysis on all books for which we have at least 100 reviews (corresponding to 50,665 reviews with 287,783 sentences for 268 books). On this selection, our impact model finds 145,768 impact expressions (2.9 per review or 0.5 per sentence). We will look at narrative and aesthetic feeling and reflection. Of these, narrative feeling is the most common (19.3 per 100 sentences), followed by aesthetic feeling (10.4) and reflection (3.5%).

Impact of genres

Zooming in on some of the more popular titles, we find that, as one would expect, expressions of aesthetic and narrative feeling are frequent in reviews of the wartime drama *Sarah's key* by Tatiana de Rosnay. An analysis of the most frequently used impact terms in these categories for this book shows *aangrijpend* ('moving', 'gripping') and *indringend* ('poignant'). Perhaps surprisingly, reviews of *The Da Vinci Code* by Dan Brown often express reflection, which may

be explained by the most frequent reflective term being 'informative'. Another book which scores high on reflection is Paolo Coelho's *The alchemist*. An author who consistently scores high on narrative feeling is thriller author Karin Slaughter.

This suggests our predictor seems to reproduce some aspects of how popular titles have been received. In order to remove some of the subjectivity inherent in picking a few titles to discuss, we also computed impact as a function of publisher-assigned genre. Among our titles with > 100 reviews we had enough information (> 20 books) for four genres: General literature Dutch, General literature translated, Literary thrillers and Thrillers. Table 5 shows for these genres the impact in the main categories. The scores are normalised around the mean, so negative numbers do not imply negative impact, only less impact than the average. The narrative feelings are clearly strongest for the (literary) thrillers, as one would expect. On the other hand, the general literature categories score higher on aesthetic feeling than the thriller genres, again conforming to expectations. For reflection, though general literature results in more reflection, as expected, at the thriller side the picture is not clear-cut. In any case the numbers seem to show that according to what the predictor makes of the reviews, readers respond to the genres as one would expect based on genre characteristics.

Table 5.
Average impact scores (subtracted mean and divided by standard deviation) for four genres.

Genre	n	Narrative feelings	Aesthetic feelings	Reflection
General literature Dutch	33	-.38	1.04	.50
General literature translated	60	-.38	.70	.33
Literary thrillers	85	.69	-.63	-.34
Thrillers	28	.57	-.61	.04

Aesthetic feeling and reflection

One interesting question that the model's predictions may illuminate is the question of the relation between aesthetic impact and reflection. In the framework presented by Koopman and Hakemulder (2015), reading literary narrative may result in two things: on the one hand (real world) empathy, mostly as a function of the role-taking that results from reading narrative, and on the other reflection on the self and ultimately self awareness and self change, as the defamiliarization that is an effect of stylistic deviation may create room for reflection.

Our impact model was formulated with narrative fiction in mind, and all books in table 5 belong to that category. Therefore, the model does not provide information about the effect of narrative versus non-narrative text. However, the table does provide some evidence for the hypothesis that literature, more than narrative text in general, can evoke aesthetic feelings and reflection in readers. This is not a new finding (see e.g. Miall and Kuiken 2002), but it is important to see it confirmed outside of the laboratory.

The data do not allow us to prove or disprove the mechanism that Koopman and Hakemulder (2015) propose (defamiliarization enabling reflection); however, we can compute correlations between the three impact factors (see table 6). In the first column, third row we note that aesthetic feeling and reflection are positively correlated at the book level. This is certainly suggestive and calls for further research.

Table 6.
Correlations between the three impact factors for all books and for literature (Dutch and translated).

Impact factors	(Pearson) Correlations	
	All books (n=268)	Literature (n=93)
Narrative - Aesthetic	-.25	.19
Narrative - Reflection	-.28	-.27
Aesthetic - Reflection	.29	.17

In this column we also note that narrative feelings correlate negatively with aesthetic feelings and with reflection. There is nothing in the Koopman and Hakemulder model that suggests narrative impact and aesthetic impact should somehow be mutually exclusive. Kuijpers (2014) argued that narrative and stylistic feeling are not necessarily in contradiction. Still, the column shows that books that are strong in narrative impact will usually be weaker in stylistic impact and reflection. It is interesting to note that the negative correlation between narrative and stylistic impact becomes positive when we restrict our attention to literary books only (second column of table 6). We surmise that narrative impact may really consist of multiple aspects, and only one of them (i.e. suspense), is negatively correlated with aesthetic feeling. Once we remove the books strongest on suspense (the thrillers) the other aspects of negative impact predominate. But this again is an issue for further research.

Discussion and prospects

We summarize what we have done, where we encounter difficulties and where we see further possibilities. We have identified four aspects of reading impact: emotional impact, narrative and aesthetic feelings and reflection. We have created a model, consisting of rules that, based on sentences from online book reviews, decides whether the sentences express one of these impact aspects. We have held a survey where frequent readers of online reviews assign these aspects to sentences from reviews, and calculated the interrater agreement for their answers. Then, for the sentences where interrater agreement was sufficient, we calculated the agreement between the raters' answers and the model's decision. In a preliminary application of the model, we looked at what the model tells us about the impact of some frequently reviewed novels. We also looked at impact by genre and found that the model at first sight seems to predict sensible

things. Moreover, it gives some indications for the existence of an often-discussed relation between aesthetic response to literature and reflection.

At various locations in the text, we have discussed limitations of our approach. First: online reviews are not necessarily immediate expressions of readers' 'true' opinions, and they need not be representative of readers in general. Second: the aspects of impact that we use are described using abstract psychological categories. It is not immediately clear how they map to literary concepts (such as style) or formulations in the reviews. This creates problems at multiple stages of our research setup: (a) The formulation of rules becomes difficult when there are no precise criteria for what constitutes e.g. reflection. All of the aspects can also be present in degrees, but is there a minimum amount of, e.g., reflection, before we recognise it as such? (b) These same issues must have arisen for the participants in our survey, some of whom have no doubt been more patient than others in following our instructions. This has led to (on average) less than satisfactory interrater agreement for narrative feeling and reflection. (c) In the comparison between the raters' scores and the model's decision, the necessarily limited accuracy of the rules and the raters' doubts come together. From that perspective it is a pleasant surprise that it is only for the reflection aspect that we do not see a significant difference in ratings between the sentences with and without predicted impact.

These limitations, in particular the not quite satisfactory inter-rater agreement scores, require us to think through once more how to take our approach further.

1. We might have provided extensive training to our subjects, in order to get them to better understand the concepts, but to some extent these are inherently fuzzy. The concepts are operationalised by the rules that we formulated, but if we would explain the concepts by showing the raters the rules, the raters would apply the rules that they were supposed to validate. Still, a training session no doubt would have helped.
2. What we might have tried is to provide the raters with the context of the sentences that they rated. This would certainly have made it easier to understand the sort of impact that a sentence expresses. However, the instructions would have become more complicated: the raters would still have to judge the specific sentence, not an entire paragraph. Our expectation is that the resulting numbers would be very hard to interpret.
3. One way of avoiding the issues of conceptual fuzziness and the difficult relation between the concept and its possible expressions would have been to use impact concepts that are closer to linguistic expressions that are actually used in the reviews. When we would, e.g., have asked raters for the presence of surprise or humour or suspense, we would probably have found much more agreement than for the more abstract impact categories that we employed. This might also make it possible to take into account the individual components of narrative feelings we mentioned above. However, it would become harder to connect our investigation with Koopman and Hakemulder's (2015) research framework. One option would be to develop the sort of instrument that we aim for, targeted at the more concrete and more directly empirical response categories, and to have these validated by our raters. The higher-level psychological constructs could perhaps be seen as combinations of the empirical response categories (e.g. narrative feeling as a combination of suspense, compellingness and perhaps other categories). The higher-level constructs could be validated not by raters in a survey of the kind we held, but by a more informal analysis of the discourse in print reviews.

4. A completely different approach would have been to dispense with the manual formulation of rules and to have applied machine-learning technology based on the ratings. The ratings would then function as a training set on the basis of which the computer would formulate the rules. This is a common procedure in the field of AI. We have not tried it because we felt that the direct contact with the text that was necessary for formulating the rules is invaluable in getting to know the language used for expressing impact. Also, machine learning would need a large amount of training data, more than we expected we could reasonably expect from our procedure. And in the case of weak interrater agreement scores the machine would have no obvious target to aim for.
5. Finally, it is also possible to argue that it was a mistake to ask 'ordinary readers' to validate what are essentially scholarly concepts. As the raters will be insufficiently aware of the scholarly background of these concepts, they will interpret them based on individual associations. Because of this, imperfect interrater agreement is only to be expected. What we should aim for instead is validation by applying our model to open research questions such as we began to do in the last section. That would be a pragmatic approach, validating the model by its capability to shed light on research issues. Its usefulness would have to be proven in practice.
6. Quite apart from considerations of interrater agreement, confronting the raters' interpretation of review sentences with the model prediction has taught us a number of things: first, there is room for improvement in the rules, especially in the area of reflection; second, maybe we should go beyond words and include punctuation in our analysis (e.g. question marks); and third, we should have been more consistent in allowing, or not, inferential rules (rules that go beyond the textual expression to infer its probable cause).

What seems to us most fruitful at this stage is a combination of points 1., 3. and 5. We will develop a model that uses smaller, more coherent and more linguistically based units of impact. These should be easier to recognize in a survey, especially when we prepare raters more thoroughly. As a second step, based on an analysis of the scholarly literature, we will assign these impact units to larger theoretical constructs such as the ones we used in this article, leading to a more principled approach than the one we have been using in this article. Meanwhile, we will continue using the model in order to test its suitability in use. There are many contexts for its application: for instance, the analysis of book blogs, understanding preferences of individual reviewers, or studying impact on the basis of the book's text (rather than that of the review). We remain convinced of the feasibility of studying reading impact in online reviews and the potential of the rule-based approach for doing so. Especially the genre-based impact differences that we found seem to us clear indications for the future possibilities of this approach.

In this contribution, we have begun to explore what online reviews can teach us about reader response. There are many ways to deepen this investigation. For instance, we have not taken into account individual differences between readers. However, for many reviewers we know gender, age and even their individual reading history. It would be a natural extension to investigate how these personal characteristics influence their response to new works. Beyond

the impact factors that we have looked at, reviews could tell us about scenes or characters that struck readers, other books or writers that they bring to bear on new works, critics or mentors that they consider relevant to their reading, aspects of the books they think are worth discussing and a host of other response-related issues. We believe that, as yet, we have only scratched the surface of the possibilities for computational research that large collections of online reviews facilitate.

References

Albrechtslund, A.-M. B. (2019). Amazon, Kindle, and Goodreads: implications for literary consumption in the digital age. *Consumption Markets & Culture*, 1-16. doi:10.1080/10253866.2019.1640216

Appel, M., Koch, E., Schreier, M., & Groeben, N. (2002). Aspekte des Leseerlebens: Skalenentwicklung. *Zeitschrift für Medienpsychologie*, 14, 149-154.

Bachmann-Stein, A. (2015). Zur Praxis des Bewertens in Laienrezensionen. In H. Kaulen & C. Gansel (Eds.), *Literaturkritik heute. Tendenzen–Traditionen–Vermittlung* (pp. 77-91). Göttingen: V&R Unipress.

Boot, P. (2011). Towards a Genre Analysis of Online Book Discussion: socializing, participation and publication in the Dutch booksphere. *Selected Papers of Internet Research*, IR 12.0. https://journals.uic.edu/ojs/index.php/spir/article/viewFile/9076/7167

Boot, P. (2012). Contextual factors in literary quality judgments: A quantitative analysis of an online writing community. Paper presented at Digital Humanities 2012. https://pure.knaw.nl/ws/files/474742/2012_boot_02_contextualfactors.pdf

Boot, P. (2017). A Database of Online Book Response and the Nature of the Literary Thriller. Paper presented at Digital Humanities 2017. https://dh2017.adho.org/abstracts/208/208.pdf

boyd, d., & Ellison, N. B. (2008). Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication* , 13(1), 210-230.

Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology*, 12(4), 321-347.

Dal Cin, S., Zanna, M. P., & Fong, G. T. (2004). Narrative persuasion and overcoming resistance. In E. S. Knowles & J. A. Linn (Eds.), *Resistance and persuasion* (pp. 175-191). Mahwah (NJ) & London: Lawrence Erlbaum Associates.

Domsch, S. (2009). Critical genres. Generic changes of literary criticism in computer-mediated communication. In J. Giltrow & D. Stein (Eds.), *Genres in the Internet: issues in the theory of genre* (pp. 221-238). Amsterdam: John Benjamins Publishing Company.

Driscoll, B., & Rehberg Sedo, D. (2019). Faraway, So Close: Seeing the Intimacy in Goodreads Reviews. *Qualitative Inquiry*, 1077800418801375. https://journals.sagepub.com/doi/abs/10.1177/1077800418801375

Berthoud, E., & Elderkin, S. (2013). *The Novel Cure: An A to Z of Literary Remedies*. Edinburgh, London: Canongate Books.

Finn, E. F. (2011). *The Social Lives of Books: Literary Networks in Contemporary American Fiction*. (PhD), Stanford University. Retrieved from https://stacks.stanford.edu/file/druid:mk148kb9574/The%20Social%20Lives%20of%20Books-augmented.pdf

Gerrig, R. J., & Rapp, D. N. (2004). Psychological processes underlying literary impact. *Poetics Today*, 25(2), 265-281.

Gruzd, A., & Rehberg Sedo, D. N. (2012). #1b1t: Investigating Reading Practices at the Turn of the Twenty-first Century. *Mémoires du livre*, 3(2). https://www.erudit.org/en/journals/memoires/2012-v3-n2-memoires0117/1009347ar.pdf

Gutjahr, P. C. (2002). No Longer Left Behind: Amazon.com, Reader-Response, and the Changing Fortunes of the Christian Novel in America. *Book History*, 5, 209-236.

Hajibayova, L. (2019). Investigation of Goodreads' reviews: Kakutanied, deceived or simply honest? *Journal of Documentation*, 75(3), 612-626. doi:10.1108/JD-07-2018-0104

Hosoya, G., Schindler, I., Beermann, U., Wagner, V., Menninghaus, W., Eid, M., et al. (2017). Mapping the conceptual domain of aesthetic emotion terms: A pile-sort study. *Psychology of Aesthetics, Creativity, and the Arts*, 11(4), 457.

Jaakkola, M. (2019). From re-viewers to me-viewers: The #Bookstagram review sphere on Instagram and the uses of the perceived platform and genre affordances. *Interactions: Studies in Communication & Culture*, 10(1-2), 91-110.

James, L. R., Demaree, R. G., and Wolf, G. (1984). Estimating within group interrater reliability with and without response bias. *J. Appl. Psychol.* 69, 85–98. doi: 10.1037/0021-9010.69.1.85

James, L. R., Demaree, R. G., and Wolf, G. (1993). rwg: an assessment of within group interrater agreement. *J. Appl. Psychol*. 78, 306–309. doi: 10.1037/0021-9010.78.2.306

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education*, 38(12), 1217-1218.

Keen, S. (2007). *Empathy and the Novel*. Oxford: Oxford University Press.

Knobloch-Westerwick, S., & Keplinger, C. (2006). Mystery appeal: Effects of uncertainty and resolution on the enjoyment of mystery. *Media Psychology*, 8(3), 193-212.

Knoop, C. A., Wagner, V., Jacobsen, T., & Menninghaus, W. (2016). Mapping the aesthetic space of literature "from below". *Poetics*, 56, 35-49.

Koopman, E. (2016). Effects of "Literariness" on Emotions and on Empathy and Reflection After Reading. *Psychology of Aesthetics, Creativity, and the Arts*, 10(1), 82-98.

Koopman, E. M. E., & Hakemulder, F. (2015). Effects of literature on empathy and self-reflection: A theoretical-empirical framework. *Journal of Literary Theory*, 9(1), 79-111.

Kuijpers, M. M. (2014). *Absorbing stories. The effects of textual devices on absorption and evaluative responses*. (Ph D), Utrecht, Utrecht.

Kuijpers, M. M., Hakemulder, F., Tan, E. S., & Doicaru, M. M. (2014). Exploring absorbing reading experiences. *Scientific Study of Literature*, 4(1).

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational research methods*, 11(4), 815-852.

Lindell, M. K., and Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Appl. Psychol. Meas*. 21, 271–278. doi: 10.1177/01466216970213006

Martens, M. (2016). *Publishers, Readers, and Digital Engagement*. London: Palgrave McMillan.

McDonald, R. (2007). *The death of the critic*. London, New York: Continuum International Publishing Group.

Mehling, G., Kellermann, A., Kellermann, H., & Rehfeldt, M. (2018). *Leserrezensionen auf amazon.de: Eine teilautomatisierte inhaltsanalytische Studie*. Bamberg : University of Bamberg Press.

Miall, D. S., & Kuiken, D. (2002). A feeling for fiction: Becoming what we behold. *Poetics*, 30(4), 221-241.

Murray, S. (2015). Charting the Digital Literary Sphere. *Contemporary Literature*, 56(2), 311-339.

Murray, S. (2018). Reading Online: Updating the State of the Discipline. *Book History*, 21(1), 370-396.

Naik, Y. (2012). Finding good reads on Goodreads. *Reference & User Services Quarterly*, 51(4), 319-323.

Naper, C. (2016). Experiencing the Social Melodrama in the Twenty-first Century. In P. M. Rothbauer, K. I. Skjerdingstad, L. E. McKechnie, & K. Oterholm (Eds.), *Plotting the Reading Experience: Theory/Practice/Politics*. Waterloo (Ontario): Wilfrid Laurier University Press.

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in health sciences education*, 15(5), 625-632.

Oatley, K. (1994). A taxonomy of the emotions of literary response and a theory of identification in fictional narrative. *Poetics*, 23(1-2), 53-74.

O'Neill, T. A. (2017). An overview of interrater agreement on Likert scales for researchers and practitioners. *Frontiers in psychology*, 8, 777.

Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. Paper presented at the Proceedings of the 21st international conference on World Wide Web, Lyon, France. https://dl.acm.org/citation.cfm?id=2187864

Papacharissi, Z. (2011). Conclusion: A networked self. In Z. Papacharissi (Ed.), *A networked self: Identity, community, and culture on social network sites* (pp. 304-318). New York: Routledge.

Pianzola, F., Rebora, S., & Lauer, G. (2020). Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers’ comments in the margins. *PloS one*, 15(1), e0226708.

Rebora, S., Lendvai, P., & Kuijpers, M. (2018). Reader experience labeling automatized: Text similarity classification of user-generated book reviews Paper presented at the EADH 2019, Galway. https://eadh2018.exordo.com/files/papers/76/final_draft/Goodreads_EADH2018_Final.pdf

Rehfeldt, M. (2017a). „Ganz große, poetische Literatur–Lesebefehl!“ Unterschiede und Gemeinsamkeiten von Amazon-Rezensionen zu U-und E-Literatur *Lesen X. 0. Rezeptionsprozesse in der digitalen Gegenwart*. Göttingen: V&R unipress.

Rehfeldt, M. (2017b). 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 und Neumann.

Ridenour, L., & Jeong, W. (2016). Leveraging the power of social reading and big data: an analysis of co-read clusters of books on Goodreads. IConference 2016 Proceedings. https://www.ideals.illinois.edu/bitstream/handle/2142/89314/Ridenour292.pdf?sequence=1

Sairio, A. (2014). 'No other reviews, no purchase, no wish list': Book reviews and community norms on Amazon.com. *Studies in Variation, Contacts and Change in English*, 15. Retrieved from http://www.helsinki.fi/varieng/series/volumes/15/sairio/

Saricks, J. G. (2005). Articulating a Book's Appeal. *Readers' advisory service in the public library*. Third Edition (pp. 40-73). Chicago: American Library Association.

Schindler, I., Hosoya, G., Menninghaus, W., Beermann, U., Wagner, V., Eid, M., et al. (2017). Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PloS one*, 12(6), e0178899.

Smith, D. (2004). Amazon reviewers brought to book. *Guardian*. Retrieved from https://www.theguardian.com/technology/2004/feb/15/books.booksnews

Spiteri, L. F., & Pecoskie, J. (2016). Affective taxonomies of the reading experience: Using user-generated reviews for readers' advisory. Proceedings of the Association for Information Science and Technology, 53(1), 1-9.

Stein, S. (2015). Laienliteraturkritik–Charakteristika und Funktionen von Laienrezensionen im Literaturbetrieb. In H. Kaulen & C. Gansel (Eds.), *Literaturkritik Heute* (pp. 59-76). Göttingen: V&R Unipress.

Steiner, A. (2010). Personal Readings and Public Texts: Book Blogs and Online Writing about Literature. *Culture unbound*, 2(28), 471-494.

Streitfeld, D. (2012, 2012-08-25). The Best Book Reviews Money Can Buy. *New York Times*. Retrieved from https://www.nytimes.com/2012/08/26/business/book-reviewers-for-hire-meet-a-demand-for-online-raves.html

Taboada, M. (2011). Stages in an online review genre. *Text & Talk-An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 31(2), 247-269.

Thelwall, M., & Kousha, K. (2016). Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, 68(4), 972-983.

Thelwall, M. (2019). Reader and author gender and genre in Goodreads. *Journal of Librarianship and Information Science*, 51(2), 403-430.

Tóth, M., & Audunson, R. (2012). Websites for booklovers as meeting places. *Library Hi Tech*, 30(4), 655-672.

Van Noord, G. (2006). At last parsing is now operational. Paper presented at the TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles. https://www.let.rug.nl/~vannoord/papers/taln.pdf.

Van Putten-Brons, S., & Boot, P. (2017). June is Dutch Literature Month!: Online Book Reviewers and Their Role in the Transmission of Dutch Literature to the English-Speaking World. In E. Brems, O. Réthelyi & T. van Kalmthout (Eds.), *Doing Double Dutch* (pp. 313-327). Leuven: Leuven University Press.

Vásquez, C. (2014). ‘Usually not one to complain but…’: Constructing identities in user-generated online reviews. *The language of social media* (pp. 65-90). London: Palgrave Macmillan.

Wallace, L. K. (2016). “My History, Finally Invented”: Nightwood and Its Publics. *QED: A Journal in GLBTQ Worldmaking*, 3(3), 71-94.

Worrall, A. (2019). “Connections above and beyond”: Information, translation, and community boundaries in LibraryThing and Goodreads. *Journal of the Association for Information Science and Technology*, 70(7), 742-753. doi:10.1002/asi.24153