Introduction

In this paper we present a method for the analysis of entity associations that real readers make in their reviews on goodreads.com, a social reading platform - and first results and insights of our analysis.

Theories of literary reception, of reading and of readers have been based on very different understandings of "reader" (cf. Willand 2014, 59-248). Most of them do not refer to real persons with books in their hands: professional readers (Dijkstra 1994), informed readers (Fish 1970, 86), model readers (Eco 1979) or even ideal readers (Schmid 2005) are instances of those approaches that lead to hermeneutics around 1800 (Schleiermacher 1838, esp. 309f.; Iser 1976). Besides low theoretical interest, another reason for neglecting the reception of real readers in literary studies is that empirical forms of reader/reading analyses are costly and time-consuming (questionnaires, interviews, peripheral physiology and eye-related measures, fMRI, etc).

This situation has changed fundamentally within the recent years. Since readers use social media to share their thoughts about the books they read, computer-supported empirical analyses of literary reviews open doors to innovative research in this field. Since computational "understanding" differs from human understanding, common questions of reception studies had to be adapted. One of them is to ask about the function of literature for real readers and society (Gymnich et al., 2005). Our approach to answer this question is the analysis of the associations triggered by literary texts. In doing so, we limit ourselves to associations concerning living real or 'living' fictional entities, such as public figures (Donald Trump) or fictional characters (Harry Potter).
Figure 1: Typical reviews with entity associations (in this case related to the book *1984* by George Orwell) on goodreads.com.

The Goodreads platform offers readers the possibility of free exchange about literary texts in a large community. 65 million members have written over 68 million reviews by 2018, whereby the reviews focus on the content and the readers’ understanding of the books and not - as with sales platforms such as Amazon - their distribution, price, etc. (Dimitrov et al., 2015).

**Data processing**

As a basis for our analyses, the reviews were stored in a local database. The database contains a data sample of 90,762 reviews of 238 books in English. 150,907 named entities were found and they refer to 6,365 individual entities. The reviews comprise a total of about 150 million tokens.

The first processing step was to clean up the reviews, e.g., the HTML tags have been removed.
To extract the entities from the reviews we used the Stanford Named Entity Recognizer (Finkel et al., 2005). The tagger also classifies the found entities and since we are looking for living entities, we only kept those from the class "PERSON".

In the next step we disambiguated the extracted entities, since a name like "Harry" may refer to many possible owners. With the help of UKB (Agirre et al., 2009) and UKB-wiki (Agirre et al., 2015), Wikipedia pages can be assigned to the entities that represent the possible entities. For this disambiguation, UKB uses the PageRank algorithm (Page et al., 1999) which evaluates documents according to their degree of linking. As soon as names like "Ron" and "Dumbledore" are mentioned in the context of "Harry", the probability increases that the reader refers to Harry Potter, Ron Weasley and Albus Dumbledore (and not to Prince Harry, Ronald McDonald and so on). Those characters origin from the novel series Harry Potter by Joanne K. Rowling and share the same context.

The Wikipedia entries are then categorized according to the ontological status of the referenced entity, i.e. whether it is a real person or a fictional character. For this purpose, the structured and thus machine-readable knowledge base DBpedia was used. Since the disambiguation provides Wikipedia entries, we can also use them to access the corresponding DBpedia entry. In addition to ontological categories, DBpedia also provides other properties that may be of interest for further analysis, such as gender or relations between entities.

The extracted data is initially stored as a table (see below) and thus allows for a flexible processing, e.g., as a network. A row of the table contains the title of the book, the disambiguated entity (link to Wikipedia page), a list of extracted entities as mentioned in the reviews, a list of review IDs to see in which reviews the name is mentioned, the number of its mentions and whether it is a fictional character or not.

<table>
<thead>
<tr>
<th>Book Title</th>
<th>Disambiguated Entity (Wikipedia link)</th>
<th>Entities as Mentioned in Reviews</th>
<th>Review ID</th>
<th>Number of Mentions</th>
<th>Fictional Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harry Potter and the Chamber ...</td>
<td>Hermione_Granger</td>
<td>hermione, hermione_grange, hermione_jean_granger</td>
<td>554404245 ... 154666862</td>
<td>333</td>
<td>True</td>
</tr>
<tr>
<td>Harry Potter and the Chamber...</td>
<td>J._K._Rowling</td>
<td>j._k._rowing, jk_rowling, joanne_rowling ...</td>
<td>376914052 ... 1123478698</td>
<td>440</td>
<td>False</td>
</tr>
<tr>
<td>Harry Potter and the Chamber ...</td>
<td>Philip_Pullman</td>
<td>philip_pullman</td>
<td>6026966884, 64654125</td>
<td>2</td>
<td>False</td>
</tr>
<tr>
<td>Harry Potter and the Chamber...</td>
<td>Harry_Potter</td>
<td>harry_potter, harry, potter</td>
<td>376914052 ... 1123478698</td>
<td>1278</td>
<td>True</td>
</tr>
</tbody>
</table>

Table 1: Example of the extracted data from Harry Potter and the Chamber of Secrets.

1 This is problematic in a few cases: e.g., if we lack Wikipedia entries for a character, if names are repeatedly misspelled or a name refers to both a fictional character and a person of public, as in a fictional novel about Napoleon (cf. Beck 2017). In the last case, the algorithm chooses the character Napoleon, because it is more closely linked with other characters of the same novel.
Findings and Discussion: Reading Data - Reading Networks

The data as described above can be analysed and visualized in many ways. For this paper, we opted for a network\(^2\) which contains two types of nodes: books and living entities that are associated to books in the reviews of those books. A book node is linked to all associations attributed to it, whereby the weight of the edge indicates the number of reviews in which a certain association appears.

\[\text{Figure 2:} \text{ Entities (purple nodes) linked to fictional books (green nodes) by green edges and to non-fictional book (orange nodes) by orange edges.}\]

\(^2\) The Gephi-network and the data described above are freely accessible: [https://tinyurl.com/EADH2018-goodreads](https://tinyurl.com/EADH2018-goodreads)
Figure 2 shows the resulting structure of the network. It was generated in Gephi using the Fruchterman-Reingold layout algorithm. While generally all nodes tend to be as far away as possible from each other, edges force them to a specific proximity. This leads to some high level observations: non-fictional books (orange) are clustered together. The same is true for some of the fictional books, e.g., books that are part of a series or saga. But it is not text characteristics that enforces this arrangement within the network. The readers’ jointly made associations with a certain group of texts constitute the clustering.

Figure 3: Cluster of the *Harry Potter* Saga, linked by a fictional character (Hermione)

Fictional characters like Hermione form a highly weighted connection with the texts of a literary saga they are part of. However, the saga is not only associated with characters of the same fictional world (cf. Veldhues, 1995). Linked actresses/actors and directors from the movies to the books as well as authoresses/authors and characters from other fictional worlds (in this case fantasy) give reason to the hypothesis that the fantasy genre triggers a very specific reception: Readers associate the content of fantasy literature *cross worlds* and *cross media* (cf. Ryan et al., 2014). Figure 4 shows multifold associations, set off by the fantasy trilogy *His Dark Materials* by Philip Pullman.
Figure 4: Cross media and cross worlds association made by readers reading fantasy literature (His Dark Materials by Philip Pullman).

In other genres, the quality of associations differs. While many of the fantasy novels in our corpus are forced into a specific closeness with each other, the same principle organizes non-fictional and canonical books: Classics are associated with entities that could be described as "classics" of the literary field (cf. Bourdieu 1993), which includes canonical authors, artists and very well known fictional characters. But there are also differences in the specific structure of the readers' associations. As the following figure exemplifies, there are only a few short edges: few entities are linked within the close proximity of a classical book, in this case The Great Gatsby. Surely, this is because classics are seldomly part of a series. But it also leads to the assumption that authors and important characters of canonical literature are frequently mentioned in other reviews. As a result, entities are located in greater distance to their actual origin. In Figure 5, F. Scott Fitzgerald appears in the upper left corner which is close to the center of the network.
Figure 5: Association structure of classics (The Great Gatsby by F. Scott Fitzgerald).

Keeping in mind that popular entities are attracted by many books (which is: associated by many readers), it becomes obvious that the layout algorithm forces those entities into the center of the network that readers think of the most while reading. An interesting finding is that those entities - besides some characters from popular culture as Batman and James Bond - are either highly canonical or bestseller authors. We assume that the canonical authors are read intensively in school. They initiate reader biographies and as such, they remain a reference point for all the books added to those reader biographies afterwards.

Figure 6: Cutout from the center of the network: high and popular culture
Results and Outlook

As a result of access to previously unimaginably large amounts of reception data, empirical reception studies have acquired a fundamentally expanding impetus; from a methodological point of view, they were previously mainly dependent on questionnaires and peripheral physiological measurements (Groeßen 1979; Baermann 1981; Funke 2003), recently supported by medical imaging methods (Christmann/Schreier 2003; Wübben 2009). Computational linguistic methods of language and corpus processing allow the analysis of very large quantities of written statements about what readers think while reading (and how they talk about it on the internet). They also enable us to empirically model reader-attributed contexts of literary texts. And thus, we gain first insights into the unanswered question with which knowledge real readers actually associate what they learn from reading.

Our findings point at the central role that canonical literature plays in a reader's mind. They lead back to the initially mentioned question of the function of literature. If it holds true that literature stabilizes and transmits the values of a society to readers (among others stated by Gymnich et al., 2005), our findings can be understood as supporting the relevance of canonical literature: The specific ways in which canonical entities as authoresses and authors of classics are associated with books show that they have a big influence on readers (cf. Kämper-van den Boogaart 2005 for the relation of canon and school education).

In the future, we plan to explore ways of grasping the contents of reader-made associations and the quality of the entities' influences. In order to so, we have to fine-tune the disambiguation algorithm to cope better with certain peculiarities arising from using Wikipedia.

LITERATURE


