CA Journal of Cultural Analytics

May 13, 2021

Putting the Sorting Hat on J.K. Rowling's Reader: A digital inquiry into the age of the implied readership of the *Harry Potter* series

Wouter Haverals, Lindsey Geybels

Wouter Haverals, Lindsey Geybels, University of Antwerp
Peer-Reviewer: J. Berenike Herrmann
Data Repository: 10.7910/DVN/S9C7NN

ABSTRACT

Compared to the large body of research into gender, race and class in children's literature, there has been little awareness of the social construction of age in this discourse. Analysing age in contemporary fiction for young readers gives insight in how present-day society models (people of) different ages, and given the decisive role that books play in shaping children's worldviews, such research contributes to our understanding of how age norms are passed on across generations. This article explores the construction of age in J.K. Rowling's *Harry Potter* in relation to the age of the implied reader. This case study provides a unique opportunity to study age, because the main characters in every volume 'grow up' together with the implied readers. This article traces the correlation between the evolutions in form and content in J.K. Rowling's *Harry Potter* series on the one hand and an evolution in the age of its implied readership on the other. After scrutinising existing guidelines pertaining to the ideal age at which to read each book, we conduct our own digital analyses on the style and topics of the texts. As well as providing insight into the evolution of these features in the Harry Potter books, this article contributes to the ongoing discussions on the reliability of readability measures and the desirability of explicit age markers on books for young readers.

The appeal of J.K. Rowling's *Harry Potter* series (1997-2007) to an audience of different ages has contributed to its unparalleled success. At the same time, an aspect that is often brought up to explain its popularity is that "Harry grew up with his readers" (e.g. Cresci, 2016). The series seems to be ageless and age-specific at the same time. In an interview, Rowling stated that she did not start writing the series with a specific audience in mind: "I wrote something that I knew I would like to read now, but I also wrote something that I knew I would like to have read at age 10".[1] *Harry Potter* was thus conceptualized as crossover literature[2] from the start, albeit with a minimum age limit of 10. With a tendency to categorisation, libraries and bookstores have attempted to refine the individual novels' implied readerships as the

series progressed. However, the suggested age labels often seem arbitrary, or they contradict one another. Furthermore, Rowling's self-proclaimed 'writing-for-all' stands in contrast with the notion of an evolving readership, which implies a specific and changing audience for each subsequent volume in the series. As a result, the implied readership of the Harry Potter series remains largely elusive.

The challenge to gain an understanding of the matter has previously been taken up by several children's literature scholars, among which Bettina Kümmerling-Meibauer, Kate Behr and Lana Whited. All three have recognized an evolution in complexity in the Harry Potter series, which they relate to the increasing age of its readers.[3] The current article aims to contribute to this investigation by introducing techniques from the field of Digital Humanities to the debate. Mainly, our focus will be on what digital text analysis is able to capture with regard to the age of the implied reader. The advantage of a digital approach lies in its ability to provide quantitative, fine-grained analyses of several aspects in multiple books, such as formal complexity and topical evolutions. The results of these computational analyses are less sensitive to the subjectivity of a researcher than results obtained by applying traditional methods such as narratological close reading. However, rather than substituting one method for another, several quantitative types of analysis in this article will be supplemented with close reading.

Starting from the observation that texts construct an image of their implied reader, the computational tools used in this article are aimed at further studying the above-mentioned evolution in complexity of the *Harry Potter* series. The term 'complexity' in this case is understood as a combination of the formal difficulty of Rowling's writing style and maturity of the topics that are covered. The first will be addressed by measuring textual complexity through a suite of readability measures, while the second will be investigated by building interpretable topic models. With some reservations, the obtained results indeed point towards evidence for an increase in complexity. We link this to an evolution in the age of the implied reader as the series progresses, while also reflecting on the limitations and validity of the computational methods used. Specifically, readability measures will be evaluated as to their potential to add to the discussion of the age of the implied reader in children's literature.

Determining implied age: diverging schemes

In literary studies, a tension exists between different narratological concepts used to refer to readership. Two concepts that are especially at odds are those of 'implied reader' and 'real reader'. However, for children's literature in particular, the distinction between both is essential, since definitions of the genre often depend on it. In *The Hidden Adult* (2008), Perry Nodelman defines children's literature as "intended for children"[4], while Seth Lerer defines it as literature that is "read by children".[5] The concept of the 'implied reader' is far from being consensual in narratology. Wolf Schmid defines it as the "image of the recipient that the author had while writing".[6] However, he acknowledges that the implied reader can have different functions in literature which leads to different understandings of the concept. The implied reader can be the "presumed addressee to whom the work is directed and whose linguistic codes, ideological norms, and aesthetic ideas must be taken into account if the work is to be understood".[7] When the author is mistaken about the norms or abilities of this addressee, this persona will not coincide with the real reader, "the flesh-and-blood person actually reading the text".[8] A second interpretation of the implied reader coined by Schmid is the "ideal recipient who understands the work".[9] No longer manifested in the mind of the author, this image of the ideal reader is created by the work itself. In this article, we will be using the term 'implied reader' according to the second function identified by Schmid because of two reasons. First, Rowling claims that she wrote the *Harry Potter* series without a specific audience in mind. Thus, there is no presumed addressee. Second, not only did the author intend her books to be read by people of all ages, the series also attracted real readers of various ages.

These two aspects, the lack of a presumed addressee and the appeal to an audience of real readers with different ages, are characteristics of "crossover literature"[10]. More so than general literature, crossover literature complicates the endeavour of researchers investigating implied readership. Moreover, what perspectives are to be included when talking about 'implied readership'? Does it solely refer to the age that the author had in mind while writing? Or does one also take into account the age labels set by publishing houses? In this respect, Beckett emphasizes the power publishers have in determining the implied audience of children's literature.[11] She points out a general tendency in the 1990s, the decade in which the first three

instalments of *Harry Potter* were published, to explicitly market book series as directed to all ages.[12] However, this did not happen with the first book, *Harry Potter and the Philosopher's Stone* (1997), as potential publishers understood it to be aimed specifically at children. Most of them rejected the manuscript because – at ca. 90,000 words – they deemed it too long to be a children's book. After several rejections, the editorial director of Bloomsbury's children's division recommended the book.[13] The initial reception further highlighted its young audience. An early review of the first book in *The Scotsman* (28 June 1997) described Rowling as "a first-rate writer for children".[14] One of the only explicit age markers found on a *Harry Potter* book is a 1998 Smarties Book prize 'sticker' – although printed – on the cover of the first paperback edition of *Harry Potter and the Chamber of Secrets* (1998), stating that the book won the Gold Award in the '9-11 Age Category'.[15] However, the crossover success of the series proved that not all power with regard to determining the age of the implied reader lies with the publisher. After Bloomsbury picked up on the enthusiasm people of different ages showed for the series, they ceased at specifically marketing the books to a young audience.

Although the original British editions of the series do not explicitly state an age range, and Rowling herself does not disclose any information about the age of her target audience, various institutions do offer age labels or ranges for the series' implied readership. These guidelines, however, do not always conform to one another. We have selected four sets of guidelines to compare in order to explore how different institutions categorize books as a method for understanding implied readership. One of them is Common Sense Media (CSM), a web portal developed to provide trustworthy information about media in general. CSM collects user-based age ratings and reviews to guide parents and teachers in their choice of entertainment for children. Furthermore, CSM provides ratings by experts based on specific content and overall guidelines informed by child development principles.[16] According to CSM, the *Harry Potter* series can be divided into three categories, lopping together books 1 to 3, marking them "for younger kids" of ages 7 to 9, and books 4 to 6, for readers aged 10 to 11 as these "books get more intense". The final book is categorised separately, for children of 12 and above because these "kids can probably handle everything J.K. Rowling sends their way".[17] CSM's categorisation is closely connected to the real reader, since it takes into account reviews of both individual adults and children who have read the books. Other institutions focus

more on the ideal recipient, the image of the reader created by the text itself, before it is received by the public.

Instead of putting a numerical age marker on children's books, contemporary English-language publishers often provide guidelines based on the presumed reading ability. For this purpose, publishing houses can resort to the well-established Lexile framework, which rates texts according to reading comprehension, ranging from 0L to 2000L.[18] These values correspond to grade levels and can therefore be converted to age ranges. Figure 1 includes the age ranges corresponding to the Lexile scores that are proposed for the *Harry Potter* books. Interestingly, although the exact Lexile values differ for the individual books, they all appear to be in line with the supposed reading abilities of fourth and fifth graders. This suggests that the entire series can be read by children aged nine to eleven.[19] These observations will be compared to our own analyses.

The country of publication makes a great difference in the determination of the age of the implied reader. First, the American publishing house of the *Harry Potter* series – Scholastic – does include an explicit age marker on one of the books. The dust jacket of the fourth novel, *Harry Potter and the Goblet of Fire* (2000) reads: 'A New York Public Library Book for the Teen Age'.[20] Scholastic further provides a categorisation of the series according to grade level on their website, which provides teaching tools.[21] Second, the Central File of Children's Books (Centraal Bestand Kinderboeken – CBK)[22] holds records for almost all Dutch fiction for young readers as well as a large collection of children's books in other languages including English, French, German and Italian. This database is unique in its kind considering its approximated total size of 400,000 titles, as well as its large collection of data on children's literature in the aforementioned languages, original and translated. Especially useful for librarians and researchers are the meta-data provided for each title, including age labels. Since it is more customary in Dutch-speaking countries to put explicit age markers on books published for children than in English-speaking countries, CBK is able to record this valuable information in their catalogue. The age-categorisation CBK attributes to the *Harry Potter* series is included in Figure 1 as well.

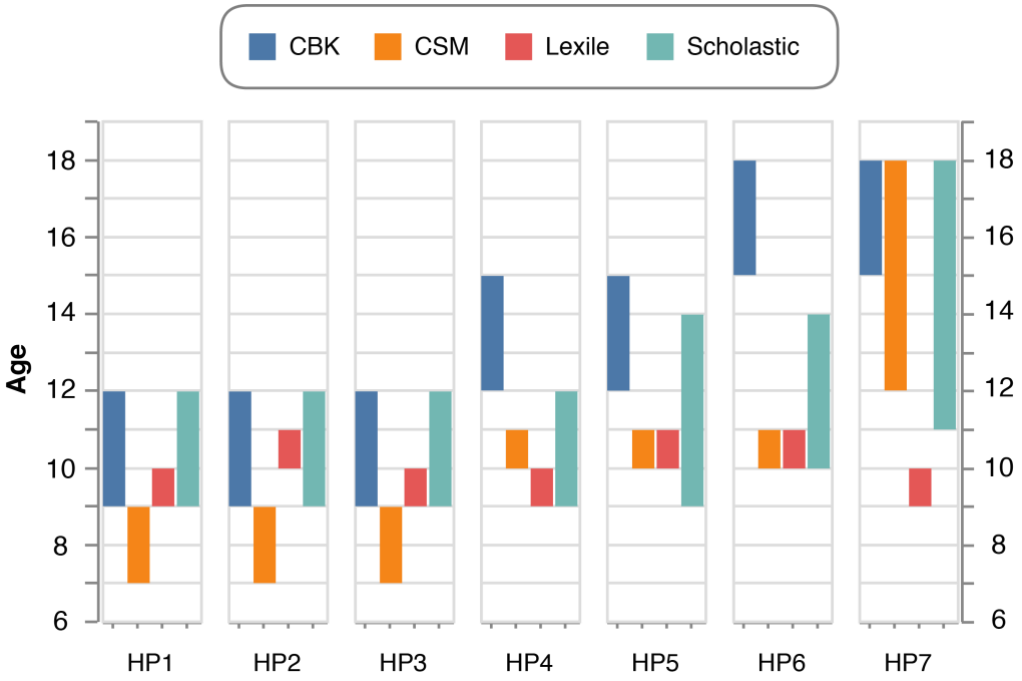

Figure 1. Age of the implied reader of the Harry Potter *series as recorded by Centraal Bestand Kinderboeken (CBK), Common Sense Media (CMS), Lexile, and the series' American publishing house, Scholastic.*

The age ranges determined by CSM, Lexile, Scholastic and the CBK show a clear overall progression, but also some differences. According to three of the four institutions, the age of the implied reader increases as the series progresses. However, they disagree on both the lower age limit and the pace at which the implied reader evolves. There is an overlap in the guidelines of Scholastic and CBK for the first three books, which are, by their standards, suitable for children aged nine to twelve. Lexile does not suggest a straightforward increase. Instead, it shows a more fluctuating trajectory between the ages of nine and eleven. Figure 1 also displays that there is a great deal of variety with regard to the age spans each authority adopts. While CBK and Lexile adhere to stable spans of respectively four and two years, CSM and Scholastic record more variable age spans. In short, not only do different authorities disagree on the numerical age recommendation for the *Harry Potter* series, their views on pace at which the age evolves and the appropriate age span also diverge.

The observations above suggest that putting a numerical age marker on children's books, or other media for that matter, is not a straightforward or unproblematic matter. The processes and necessity to do so are the subject of much debate (see Nikolajeva 1996, De Vriend 2006, Fastenau 2014). Stephen Krashen (2001) even

considers the Lexile measure "unnecessary and potentially harmful."[23] The remainder of this article will acknowledge the variation in the age ranges provided by the four institutions while adding our own, computationally-aided point of view to the discussion. What can be concluded, however, is that most guidelines agree that the age of the implied readership of the books increases as the series unfolds. The implied reader moves from childhood through adolescence and, according to two schemes, towards adulthood.

The remainder of this article explores what computational tools designed to assess textual difficulty can add to the question of *Harry Potter*'s evolving implied readership. Because we have no information on the presumed addressee as imagined by the author, the term 'implied reader' will henceforth be used to refer to the ideal recipient of the work as determined by the text itself. The aim is not to determine absolute age ranges for the series' implied readership, but rather to investigate whether the evolution suggested by the guidelines above is reflected in computationally-aided analyses of readability and topics. Moreover, by comparing several ways in which this type of analysis can be conducted, this article adds to the critical reflection on the validity of readability measures.

Determining implied age: digital analyses

The correlation of the age of the implied reader and readability of children's literature has already received scholarly interest, often in connection with literacy and education (see Meyer 1975, Fry 2002, Yi Ma and Loftus 2012). In this section, analyses are aimed at answering the following question: Do the *Harry Potter* books become more difficult to read in terms of syntax and semantics? [24] The selection of available digital analyses we made is informed by a paper by Wanner et al. (2011) in which the writers present and evaluate a tool for assessing age suitability of books.[25] Their tool combines story complexity, emotions, physical aspects, difficulty of writing style and topics. This article will focus on the last two. First, we look into formal features to examine the complexity of the *Harry Potter* series. In this respect, the average sentence length of each book is calculated as well as the number of subordinate clauses as a syntactic base for textual difficulty.[26] Next, lexical diversity will be studied by resorting to Moving Average Type-Token Ratio (MATTR). To conclude the formal analyses, this article will compare several readability formulas,

or measures that aspire to determine the minimum reading level needed to comprehend a text. In the last section of the article, previous analyses will be complemented by an exploration of the content of the *Harry Potter* series by use of topic modelling. The ultimate goal is thus to trace a possible connection between formal and content-related evolutions on the one hand and the increasing age of the implied reader on the other.

Form in the series

Sentences

We first look for an evolution of the formal features in the series. Doing so, we start by looking at sentence length which, according to Colleen Lennon and Hal Burdick, is "the best predictor of the difficulty of a sentence", and by extension of a text.[27] The average sentence length of the *Harry Potter* series as a whole is slightly less than twelve words (11.97 to be precise, s=1.16).[28] This is almost one word less than the average sentence length of J.K. Rowling's fiction written for an adult readership (12.78, s=0.61). An examination of the individual books (Figure 2) indicates a slight shift in the average sentence length between the first three volumes in the *Harry Potter* series on the one hand, and the series' subsequent volumes on the other. In order to potentially explain these differences, we examined the ratio between character speech and narration[29] and found that there is less direct speech in the first three volumes (averaging 37%) when compared to the rest of the series (averaging 39%). However, calculating the average sentence length of both types (character speech/narration) for the entire series showed that narration passages have on average 5.5 more words per sentence. The ratio of direct to indirect speech does not influence the average sentence length of the books.

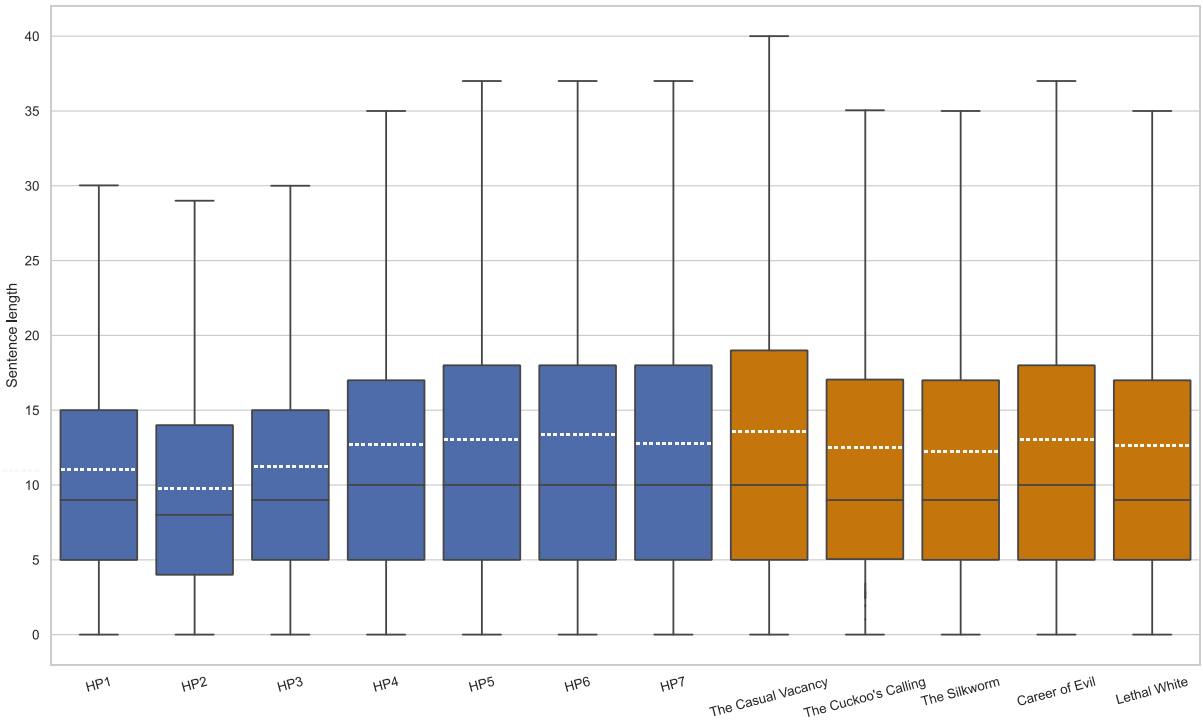

Figure 2. Sentence length of each book in the Harry Potter *series and J.K. Rowling's adult fiction (ordered by publication date). Average sentence length is represented by a white dotted line.*

Although sentence length is an important factor in assessing the difficulty of a text, it would be imprudent to consider it as a straightforward reference point for the age of the implied reader. After all, sentence length and text complexity are not necessarily directly proportional.[30] The underlying syntactic structures of longer sentences are much more indicative. Bailin and Grafstein point out that sentences with a deeper syntactic structure are more complex.[31] Especially when a subordinate clause is nested within another subordinate clause, this adds to a sentence's complexity, and thus requires stronger reading skills. Figure 3 shows for each book the proportion of sentences with at least one and more than two subordinate clauses.[32] As the series progresses, the number of sentences with at least one subordinate clause increases. These types of sentences were found to be positively correlated with the series' progression (*Pearson's r*=.83, *p*=.02). Particularly striking is the gap between the two extremes: the proportion of sentences with at least one subordinating clause in HP2 is 37.87%, whereas in HP6 this ratio lies at 63.19%. An increase in formal difficulty can also be observed in the number of sentences which contain two or more subordinate clauses. Between these more complex sentences

and the progression of the series, a positive correlation was also found (r=.81, p=.03).

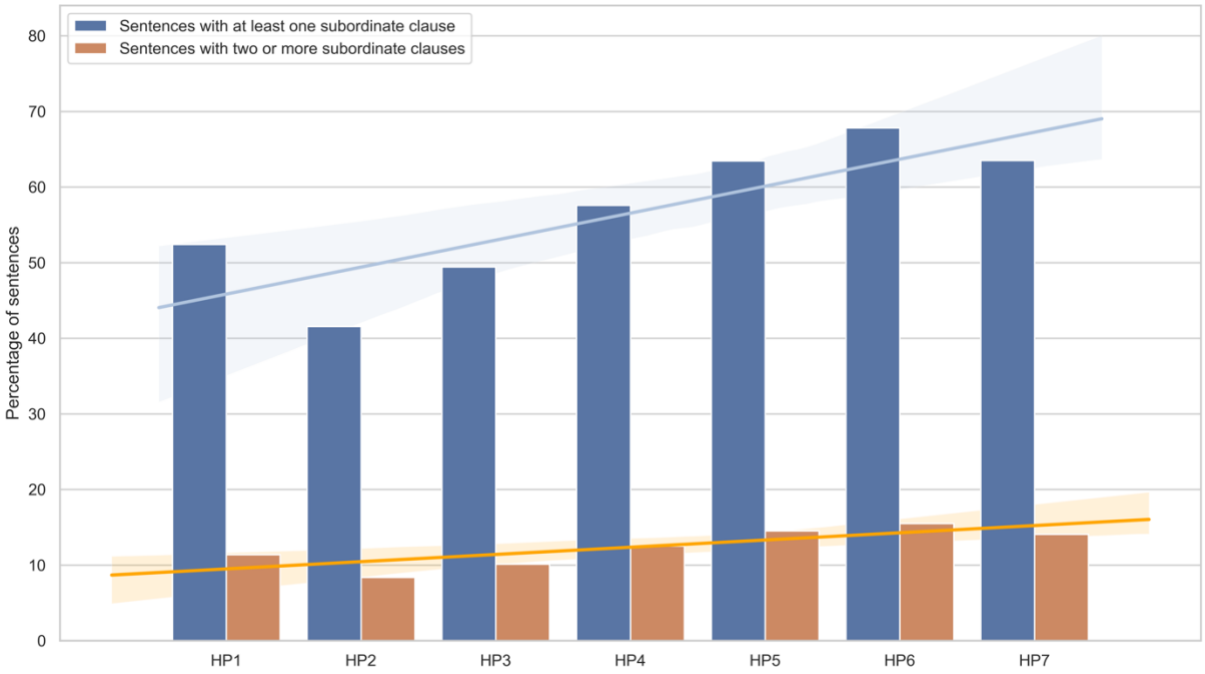

Figure 3. Ratio of sentences per novel that contains subordinate clauses. Syntactic structure was analysed using the Berkley Neural Parser *(Kitaev, Cao & Klein 2018).*

Lexical diversity

Another formal feature that influences the complexity of a text is lexical diversity, a textual feature identified by Victoria Johansson to successfully detect differences between readers in different age groups.[33] Lexical diversity represents the vocabulary richness of a text, most frequently measured by the so-called type-token ratio (TTR), which refers to the ratio of unique words – types – to the total number of words – tokens – in a text.[34] TTR outputs a number between 0 and 1; the higher the number, the more types a particular text contains. However, a notorious shortcoming of TTR is its susceptibility to text length (i.a. McCarthy & Jarvis 2007). The more words a book contains, the more previously used words will reappear, lowering the TTR-score. Therefore, instead of TTR, we picked a measure that serves our goal: determining the lexical diversity of the vocabulary of the individual *Harry Potter* books, regardless of their unequal lengths.[35] Ideally suited for this purpose is Moving Average Type-Token Ratio (MATTR), developed by Covington and McFall (2010). It is calculated by taking the average TTR value for overlapping

segments (with a fixed length) of a text. For our calculation we set the segment length to 10000 words.[36] Figure 4 shows the MATTRs per book for the *Harry Potter* series and Rowling's novels for adults. Noticeably, the ratios are unaffected by the lengths of the books: the second shortest book in the series (HP2, 85071 tokens without punctuation) has the highest MATTR (.209), whereas the lowest score (.188) is recorded for the shortest book (HP1, 76440 tokens without punctuation). As the differences in the MATTR-scores for the books are slight, it is premature to link these results to an advancement of the age of implied reader of the series. Notwithstanding this result, a remarkable observation can be made when comparing the MATTR-scores of the *Harry Potter* series (mean at .20 ± .001) to those of the novels written for an adult audience (mean at .24 ± .001). From Figure 4, we learn that MATTR-scores for the adult novels are consistently above the overall average (.22 ± .0007). Lexical diversity is relatively stable throughout the *Harry Potter* series; no correlation was found between MATTR and the order in which the books appeared. Thus, in terms of lexical diversity no evolution in complexity was found.

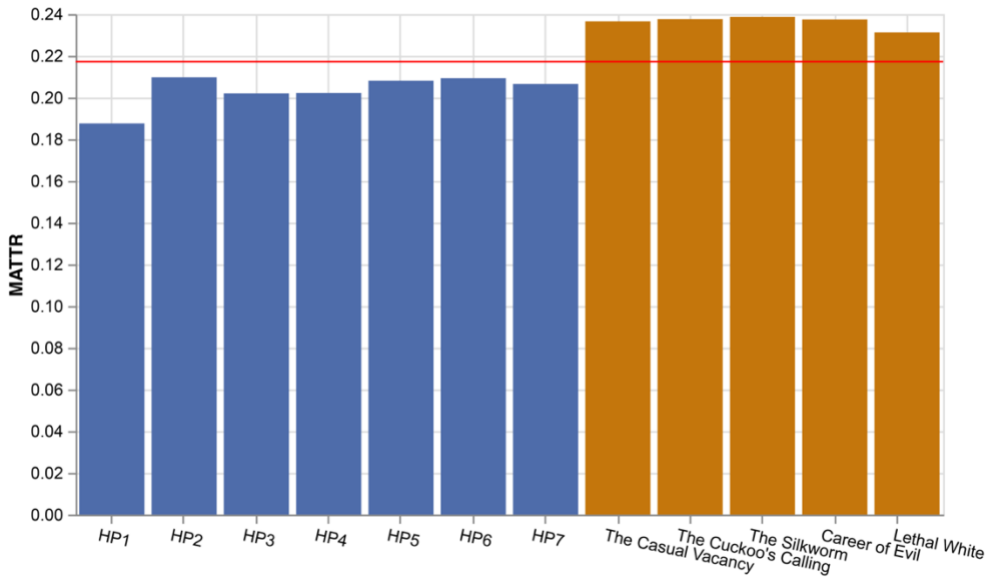

Figure 4. MATTR scores for the individual books in the Harry Potter *series and Rowling's adult novels. Calculated using a 10000-word window size. The novels appearing on the x-axis are ordered chronologically. The red line shows the mean (.22).*

Readability

Although the analyses above reveal aspects of the evolution of the books' complexity, it remains challenging to interpret them in terms of the age of the implied reader. Formulas that *do* aspire to such a correlation are referred to as 'readability formulas'. They aim to provide an estimate of the minimum reading skills required to understand a particular text. In their tool to assess age suitability, Wanner et al. (2011) include the calculation the Automated Readability Index (ARI). This readability test was developed by R.J. Senter and E.A. Smith in 1967 and combines information on the number of characters, words and sentences in a text.[37] However, Wanner et al.'s choice to integrate specifically ARI remains unsubstantiated. After all, there exist numerous, well-established readability formulas, which have been broadly applied to literature. Especially teachers have relied on readability formulas for decades to analyse children's literature and textbooks.[38] Librarians use them to aid visitors in their search for suitable reading materials.[39] Crossley et al. (2019) note that the wide use of these 'classic' formulas contrasts with the limitations in their function to determine reading levels partly due to their lack of construct validity and because they seem to be less accurate on data other than the data they were trained on. In his discussion of readability formulas, he refers to a study he conducted in 2017 in which he demonstrated the benefit of readability formulas that use "features that measure lexical and syntactic constructs, text cohesion, sentiment, topic analysis and semantics".[40]

In order to obtain a more comprehensive understanding of how different readability formulas assess the age of the implied reader, we will not limit ourselves to the application of just one formula. Rather, we chose to expand our analysis of the readability of the *Harry Potter* series with five more formulas. This enables us to compare individual formulas and evaluate the use of readability scores to study children's literature. Next to the ARI, other popular readability measures are Gunning fog, Dale-Chall, the Simple Measure of Gobbledygook (SMOG), the Coleman-Liau Index and the Flesch-Kincaid formula.[41] These formulas output a value that corresponds to the reading abilities of a student within the U.S. grade level system. Table 1 provides a conversion chart for these grades to the respective ages of students.

Grade	Age	Grade	Age
3	8-9	8	13-14
4	9-10	9	14-15
5	10-11	10	15-16
6	11-12	11	16-17
7	12-13	12	17-18

Table 1. Conversion chart for U.S. grade levels and student ages.

Since the above-mentioned measures offer a purely formal analysis of readability (rather than taking into account the content or thematic aspects or any form of empirical analysis such as reading speed), they are often heavily criticized.[42] It should be stressed, though, that in the current study readability measures are not being used to cast in stone a text's suitability for a particular age group. After all, we are well aware that a novel's readability does not solely depend on formal features. Rather, we aim to investigate whether the established readability formulas are sensitive to a potential evolution with regard to the complexity of the *Harry Potter* series, thus reflecting a tendency for a shift in the age of the implied reader.

Figure 5 shows the scores obtained for the above-mentioned readability formulas directly translated to the U.S. grade scale.[43] For this purpose, each novel was divided into samples of 200 consecutive sentences. Next, 25 samples were selected at random for which the readability scores were calculated. From a general outlook, it appears that readability scores gradually rise as the series progresses. Based on the results, both the SMOG score and Gunning fog correlate most strongly to the publication chronology (Kendall's =.39, $p<.001$). Flesch-Kincaid ($\tau=.31$), ARI ($\tau=.28$) and Coleman-Liau ($\tau=.25$) all exhibit moderate positive correlations ($p<.001$). It should be noted, though, that the aforementioned readability formulas also correlate very strongly with each other (τ ranges from .65 to .88, $p<.001$). This should come as no surprise as most formulas exploit the same textual features to arrive at a result (such as average sentence length, word counts, word length, etc.). Only the Dale-Chall formula correlates less strongly with the other formulas (τ ranges from .15 to .34, $p<.001$). Unlike the other formulas, the Dale-Chall formula

is unique because of its use of a word list containing ca. 3000 words recognized by 80% of fourth graders. This way, the number of difficult words in a passage (i.e. words that are not on the list) is factored into the formula, making it more advanced and more predictive of readability than formulas resorting to word length (e.g. SMOG, Gunning fog, Flesch-Kincaid).[44] As the Dale-Chall formula is vocabulary-based, it closely resembles the Lexile measure, created and owned by the company MetaMetrics and used by publishing houses including Scholastic to provide their books with readability scores.

From the averages in Figure 5 (white dashed line), we learn that the first book is scored as the most readable across all formulas. This result runs parallel to the analysis of MATTR scores (Figure 4). The highest average readability scores are recorded for HP6 and HP5, followed by HP7. One factor that might influence the slight decrease in readability from the sixth book to the final book might be the ratio between character speech and narration. There is less direct speech in the last book (39%) when compared to the previous one (44.5%). As established in our analysis of sentence length, narration has on average longer and more complex sentences than character speech. Almost all readability formulas identify a rise in readability between books one and six (except for Dale-Chall).

Remarkably, the grade level estimates produced by the different formulas are in some cases far apart. Most noticeably, this can be observed for Flesch-Kincaid and Dale-Chall. While Flesch-Kincaid suggests that HP1 is suitable for third graders (ages 8 to 9), Dale-Chall sets the readability level at the eighth grade (ages 13 to 14). It is likely that this is caused by the variables used in both formulas. While Flesch-Kincaid takes into account syllable, word and sentence counts, Dale-Chall is based on word counts, syllable counts, and the ratio of difficult words as recorded in a list. The observed variation between readability measures is also what sparks criticism.[45] While these variations make it undesirable to use readability measures to put reliable grade levels on the books, there is relative agreement about the directionality of the increase of complexity. This supports research pointing out that sentence length and word difficulty are viable features for estimating textual difficulty, even though imperfect.[46]

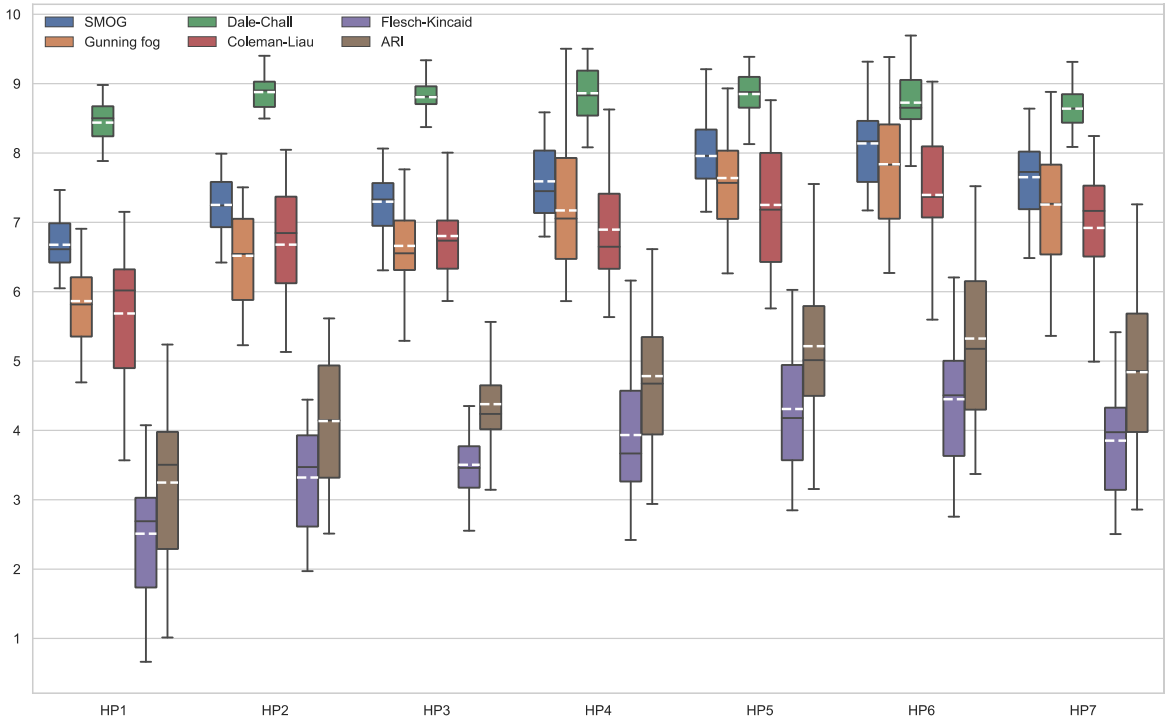

Figure 5. Readability formulas applied to 25 200-sentence samples taken from the individual books in the Harry Potter series.

Similar discrepancies can also be identified if we look at the age attributions set by recognised institutions. Table 2 provides an overview of the recommended age ranges for each *Harry Potter* book as well as the age ranges determined by the readability scores. From this table, we learn that while age recommendations by the institutions don't match up perfectly with those suggested by readability formulas, both indicate an increase in the age of the implied reader. A drawback in this respect is the intent of readability formulas to target a single, specific grade level, while the institutions often suggest age recommendations spanning more than one grade. From Figure 5, we also learn that, interestingly, all readability scores drop for the final book in the Harry Potter series. However, only the Lexile framework shows a decrease in the age of the implied reader. All other organisations report an increase. This suggests that the recommendations of CSM, Scholastic and CBK are not based solely on formal analyses and readability of the texts, as suggested in their introduction above. Although the organisations themselves are not clear about this, it would appear that, in addition to formal characteristics, content and/or thematic elements are also taken into account. To further investigate the discrepancy for HP7 between our analyses and the four institutions, but also to complement the strictly

formal analysis by a semantic one we will assess the content of the books in the next section.

	HP1	HP2	HP3	HP4	HP5	HP6	HP7
CSM	7-9			10-11			12-15
LEXILE	9-10	10-11		9-10	10-11		9-10
SCHOLASTIC	9-12				9-14	10-14	11-18
CBK	9-12			12-15		15-18	
FLESCH-KINCAID	8-9			9-10			
ARI	8-9	9-10		10-11			
COLEMAN-LIAU	11-12			12-13			
GUNNING FOG	11-12		12-13		13-14		12-13
SMOG	12-13			13-14			
DALE-CHALL	13-14			14-15			
ALL*	10-11	11-12		12-13			

Table 2. Age ranges for each Harry Potter *book corresponding to the examined readability formulas. *The final row ('All') contains the corresponding age ranges if we were to aggregate all the readability scores across all formulas.*

From childlike to mature topics

After analysing the evolution of formal characteristics of the *Harry Potter* series, we look at topics which decrease or increase over the course of the series. For these analyses, topic modelling is used.[47] The analysis of topics is one of the components Crossley et al. (2019) propose to include in new, improved readability formulas.

Topic modelling fits into the field of distributional semantics and is thus concerned with the subject matter of documents, examining *what* is written in a text as opposed to *how* it is written. Topic models generate clusters of words that frequently appear together (word co-occurrence). The meaning of a word can be approximated by looking at its context. First the topic model is "trained" on a large corpus, in this case The Books Corpus,[48] to identify word clusters.[49] The number of topics identified by the model has a strong influence on the results. To accommodate this variation, the model was trained three times; with 100, 200 and 300 allowed topics. Next, these word clusters, or topics, were tested on the *Harry Potter* series, divided into their original chapters and pre-processed to retain only content words, to see to what extent each topic is present.[50] In practice, this means that 199 chapters each received the same number of scores as there are topics trained in the model.

To identify the topics that increase or decrease most significantly throughout the series, the Kendall rank correlation coefficient, Kendall's Tau, is calculated for each topic. This statistical test represents how consistently a score decreases or increases. A set of scores with no clear pattern, and thus not interesting to our analyses, will not receive a significant Kendall's Tau. The result from applying the statistical test is a list of top increasing or decreasing topics. It is important to note that topic models only create clusters of words that are semantically closely related; reliably labelling these clusters in most cases requires human input. Consequently, the interpretative phase of topic modelling is more prone to bias than the analytical phase. To minimise a subjective reading of the topics provided by the Kendall's Tau test,

Table *3* presents the results of all three versions of the topic model. Five words are given as examples for each topic. These are not necessarily words that are present in the *Harry Potter* books, but rather the topmost characteristic words for the topics based on the background corpus. This is clear in topic 6 of the most decreasing out of 100 topics, which can be attribute to the large amount of fan fiction included in the Books Corpus.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Decrease (100 Topics)	dog	**game**	class	doctor	kitchen	shit
	dogs	team	**school**	**hospital**	bathroom	gon
	puppy	ball	teacher	nurse	bedroom	ass
	cat	play	students	patient	shower	whispered
	animals	players	classroom	patients	stairs	babe
Decrease (200 Topics)	letter	**school**	class	cat	**game**	dog
	letters	year	teacher	cats	team	puppy
	envelope	college	students	**animals**	ball	pet
	read	summer	**classroom**	animal	play	**animal**
	paper	grade	classes	shelter	players	tail
Decrease (300 Topics)	**dog**	**food**	class	**game**	**school**	**candy**
	pet	eat	classes	ball	football	chocolate
	tail	plate	**classroom**	play	grade	tickets
	barking	eating	lunch	players	teachers	cookies
	fur	meal	today	baseball	college	cookie
Increase (100 Topics)	**wedding**	soldiers	tears	replied	wife	church
	dress	soldier	loved	answered	husband	altar

	marry	**military**	whispered	exclaimed	daughter	soul
	married	tent	cry	explained	married	**souls**
	bride	riffle	crying	shouted	**marriage**	angels
Increase (200 Topics)	**death**	army	truth	sword	daughter	mage
	died	**war**	**conversation**	blade	husband	ivory
	die	battle	feelings	**swords**	daughters	silver
	killed	enemy	tone	hilt	birth	pack
	alive	troops	trust	dagger	age	healing
Increase (300 Topics)	**ring**	**death**	answered	tones	nature	daughter
	finger	died	stated	gentleman	**social**	daughters
	diamond	funeral	explained	demanded	order	birth
	engagement	die	group	**fear**	course	age
	rings	grave	responded	features	knowledge	parent

Table 3. Top six increasing and decreasing topics in the Harry Potter *series as identified by a topic model trained to distinguish between 100, 200 and 300 topics.*

When looking at the results of the top six decreasing topics, all three models identify topics related to animals, sports and school. The presence of animals in fiction is traditionally connected to literature for children. In her discussion on fictive characters, Maria Nikolajeva (2002) connects nonhuman characters, including animals, to childhood.[51] Animals play a larger part in the first books because of the

introduction and diversity of magical creatures to which less attention is paid in the later volumes. For example, an important part of first-year students' school experience is choosing a pet. In Figure 6 this thematization can be observed. The graph is a visualisation of the Kendall's Tau slope resulting from the topic model trained on 300 topics. The graph is plotted using a rolling window of 35 chapters to stabilise the scores of each topic; the graph starts in the middle of this window, at chapter 17, which corresponds to the beginning of the second book and thus no data is shown for HP1. The trend line of the topic on dogs (dog pet tail barking fur) is high at the start of HP2, most probably influenced by the presence of a three-headed dog at the end of HP1, and peaks in HP3 due to the thematization of the grim figure of a black dog. According to Behr (2005), the evolution of the topic of magical details, such as magical creatures and animals, is closely linked to the age of the implied reader.[52] While young readers are drawn to these details of the marvellous wizarding world, the later books lack these topics.

Competitive sports and games are also often mentioned in the first part of the series and are represented in all three models. Parallel to the importance of magical creatures, magical games and sports are thematised as part of Harry's introduction into the wizarding world. Figure 6 shows a high presence of this topic at the start of books two and four. From a close reading we learn that the first chapters of the fourth book are set at the Quidditch World Championship. While Quidditch, the magical team sport played at the school, continues to feature throughout the series, with the exception of the last book, the analysis points towards a decrease in the presence of this and similar activities that feature a clear divide between teams. One reason for this decrease can be found in Jann Lacoss' (2002) observation that in the series the separation into groups, such as Quidditch teams but also schoolhouses, is more defined in the first books while more mixing occurs in the later books.[53] We can hypothesise from this that group membership is more important and straightforward for children than it is for adolescents. In later volumes, the topics of sports and animals are not featured as much in part because the novelty of these magical elements has worn off and partly because they are overshadowed by more serious tasks at hand.

All three models present a decrease in the topic of the school setting. The topic model trained to identify 200 topics even includes two topics related to this evolution in the

top six of decreasing topics. According to Nikolajeva (2010), the school setting in the *Harry Potter* books emphasises the power structure of adult over child.[54] The decrease of this topic as the series progresses indicates a moving away from childhood and towards adulthood. Figure 6 suggests that this movement is halted in books five and six, where the presence of the school topic increases before dropping again in the last book. Based on a close reading of the books, this is an accurate rendition of the topic, as books five and six focus more on magical education than the previous and last books.

Furthermore, various scholars (e.g. Nikolajeva 2002) have drawn a parallel between sexuality in general fiction and food in fiction for children. Food is also one of the topics decreasing in importance. While it only turns up in one of the models, the one trained to distinguish between 300 topics, it does so twice in the top six of said model, once centred around the act of eating and once concerned more with sweets. The decrease in the topics concerning food, as well as animals, sports and the school setting, points towards a maturation of the series and consequently an evolution in its implied readership.

Figure 6. Evolution of the five topics with the steadiest decrease in the Harry Potter *series.*

The same conclusion can be drawn from examining the topics that increase in the *Harry Potter* series. Although there is more variation between the three trained models, most topics they identified are linked to a more adolescent or adult experience. Two models present topics on battle and war; it is also present in the analysis of the most detailed model (300 topics), but there it falls just short of the top six. According to J.A. Appleyard (1994) children's literature is no place for war and violence. Whereas he observes that good and evil are not clearly separated in adolescent literature, in children's literature evil is externalized and overcome.[55] Although exceptions to both of Appleyard's findings can be found in contemporary children's literature, Rowling's series is initially set up to comply with the traditional convention of a fairly innocent world in which good and evil are distinguished, and that this world gradually grows more complex. It is true that in the last *Harry Potter* book the personification of evil is defeated, but it does not happen without several losses on the side of good. Another theme that is often featured in adolescent literature is fear. Behr states that feelings of wonder and innocence make way for fear and tragedy in the *Harry Potter* series,[56] effectively connecting the decreasing topics on the magical details of the wizarding world to the increase in the topics of fear, evil and death. This last theme was also identified as increasing significantly by two of the topic models and validates the claims of several literature scholars studying the *Harry Potter* series that death is one of the main themes of the books (see i.a. Trites 2001, Cockrell 2002, Behr 2005). Figure 7 shows the topics with the steadiest increase in the series. Only the top three topics have a large enough increase to create a meaningful plot. The topic of death has two peaks in the course of the series: one at the end of HP4 and one beginning in HP6, reaching its highest point at the end of the series. The first peak correlates with arguably the first major death in the series and the moment that evil is reborn.

Clearly present in the analyses of increasing topics is the topic of family setting. All three models include in their top six words such as daughter, husband, wife, birth, age, and two of them include words associated with marriage. Although there is indeed a wedding at the start of the last book, the models probably pick up on the importance of a magic ring featured in the last two books, as observed in one of the topics of the most detailed model. The family setting is traditionally more associated with children's literature, as adolescent characters are often depicted as rebelling against their family. In the sphere of social relations, there is one topic lacking in the

later *Harry Potter* books which we would expect to see in adolescent literature, that of romantic relationships and sex. Children's and YA literature scholars including Appleyard (1994) and Lee Talley (2011) recognise sex as one of the main differences between both types of fiction. While children are usually shielded from this topic, it is present in most adolescent literature. The parallel that Nikolajeva (2002) draws between sexuality in general fiction and food in fiction for children[57] is not visible in the topic model. While topics on food have already been shown to decrease, the topics included in

Table *3* do not support a complementary increase in sexuality with regards to the *Harry Potter* series.

One more interesting observation from the increasing topics is the presence of topics about conversation. One model has a topic on truth and feelings, which also reflects Appleyard's observation of adolescent literature often featuring the "turbulent emotions of the central characters".[58] The other two models identify topics characterised by dialogue tags (replied, exclaimed, stated, etc.). A possible explanation for the increase in words like these is the action-oriented nature of children's fiction, while dialogue, reflection and description are associated more with adolescent or general fiction.[59] A different explanation might be that this is an artefact of the evolution in writing style of the author, that instead of using the generic 'said', Rowling's description of characters' speech became more diverse.

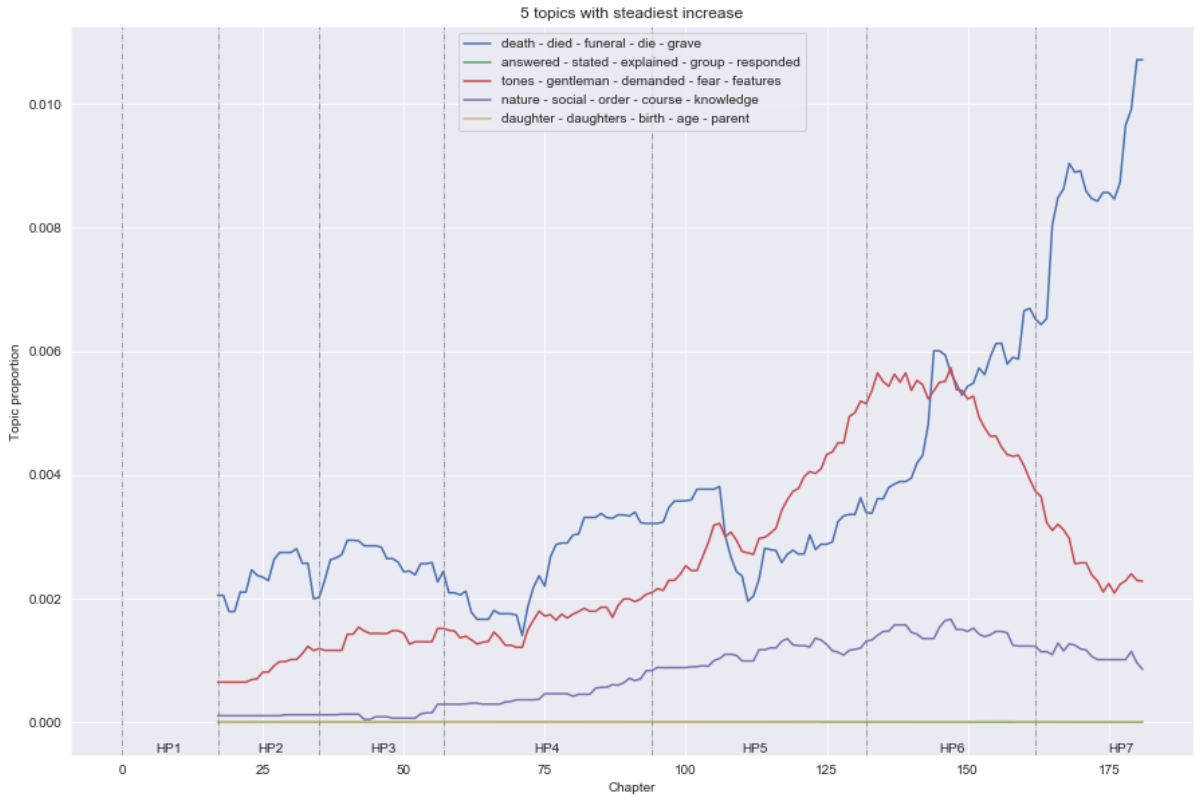

Figure 7. Evolution of the five topics with the steadiest increase in the Harry Potter *series; only the top three show a meaningful evolution.*

Conclusion: maturing text, ageing readers

One of the main goals of this article was to trace a possible correlation between the evolution in the complexity in form and content of the *Harry Potter* series on the one hand to the evolution in the age of its implied readership according to various institutions on the other. Firstly, we established that it is problematic to assign reading age to the individual *Harry Potter* books – because of their crossover nature and the refusal by both the publisher and the author to make explicit assertions. However, three of the four institutions discussed in this article agree with literary reviewers and scholars and recognise an evolution in the age of the implied readership. In the second part of this article, we investigated whether this evolution can be picked up by a digital analysis of the texts. Although the *Harry Potter* series was apparently written without a specific audience in mind and it was quickly marketed to be suitable for all ages, the analyses conducted in this article were able to profile to a certain extent an age-dependent implied reader. Both the formal and topical analyses show a change throughout the *Harry Potter* series.

When looking at formal aspects, the average sentence length and number of subordinate clauses show a rise in difficulty. The increase in lexical diversity is slight, and lower than that for the novels written for an adult audience. While the readability formulas show some variation between each other (e.g. a four-year difference between the ranges determined by Flesh-Kincaid and SMOG for almost each book), they all support the assumption of increasing difficulty across the series reported by literature scholars and reviewers as well as the institutional guidelines discussed. However, the measures indicate only very small changes. While previous studies into reading abilities of children and English-language learners prove readability measures to be valuable (especially in educational contexts), this article complements other research that identifies the concept of 'readability' as being too complex to infer conclusions from the analysis of only one aspect. Adding complementary formal as well as topical analyses is therefore necessary to get a richer image of the implied reader. However, while we established the utility of these analyses on a specific corpus, we remain well aware of the discussion on their desirability. Further research, conducted on a larger corpus of children's literature that does include age markers made explicit by the author or publisher, would possibly provide a more detailed insight into the validity of the methods employed in this article to determine a correlation between formal and topical features on the one hand and the age of the implied reader on the other. Especially interesting would be to apply these computational tools to the oeuvre of crosswriters who write for adults as well as children of different ages.

While most of our analyses pick up on a general evolution in complexity in the *Harry Potter* series, some results raise questions pertaining to the validity of the computational tools used. Especially the contradicting results of the analyses conducted on HP2 signal a problem; while there is a decrease in sentence length and number of sentences containing subordinate clauses, a higher lexical diversity is recorded as well as a lack of decrease in the age guidelines either of institutions or as calculated by readability formulas. The reliability of using topic modelling to study the age of the implied reader is also debatable. Linking topics to an evolution as detailed as the small age ranges suggested by the existing schemes is challenging as it is difficult to connect the presence of certain topics to a narrow age range. There is no measure for example to determine how much talk about death a reader of a certain age can deal with. It is also important to note, as illustrated by the presence

of the keyword 'ring' and its possible influence on the topic of marriage, that a close reading of the texts remains valuable when employing digital tools. Nonetheless, if we match the evolution of topics with critics' discussion of children's and (young) adult literature, it is clear that there is a general movement from childhood topics to adolescent or even adult topics. The decrease of topics concerning food, school and animals combined with the increase in spiritual and morbid themes point to a maturing of the content of the series. The sudden rise in the topic on 'death' in the last two books might suggest a change in implied readership between these and the previous books. Butler (2003) recognises an increase in the age, reading levels and maturity levels of readers as a consequence of the maturing of themes as well as of fictional characters.[60]

Notes

[1] Quoted in: Philip Nel, *J.K. Rowling's Harry Potter Novels: A Reader's Guide* (A&C Black, 2001): 51.

[2] Sandra L. Beckett, *Crossover Fiction: Global and Historical Perspectives* (New York: Routledge, 2009).

[3] The exact nature of this relation has been subject to a debate that lies beyond the scope of this article. Bettina Kümmerling-Meibauer ('Seriality in Children's Literature.' In Clémentine Beauvais and Maria Nikolajeva (Eds.), *The Edinburgh Companion to Children's Literature* (2017): 167-178.) attributes the increase in complexity to the maturation of the title-character, which she links to the ageing reader, as both effectively 'grow up' together. Other researchers draw a more direct connection between the changing readership of the series and its evolution but do not agree on the direction of this causal relationship; Sandra Beckett (2009) and Victor Watson ('Series Fiction.' In Peter Hunt (Ed.), *International Companion Encyclopedia of Children's Literature* (2004): 532-541.) believe that the series' growing success with adult readers and the original child readers growing up led to an evolution in content while Rebecca Butler (2003) states that the change in content precedes the change in readership.

[4] Nodelman, Perry. *The Hidden Adult: Defining Children's Literature* (Baltimore: Johns Hopkins University Press, 2008): 20. Further complicating the matter, is the fact that the intended audience of children's books is traditionally determined by adult mediators involved in their production and distribution to the child reader. Mostly, the authors' involvement in this decision is limited since age markers are usually discussed with or even imposed upon their work by publishers, booksellers, literary critics or librarians.

[5] Seth Lerer, *Children's Literature: A Reader's History, from Aesop to Harry Potter* (Chicago: University of Chicago Press, 2009): 11.

[6] Wolf Schmid, "Implied Reader," *The Living Handbook of Narratology* (2014). https://www.lhn.uni-hamburg.de/node/59.html.

[7] Ibid.

[8] Gerald Prince, "Reader," Peter Hühn et al. (Eds.) *Handbook of Narratology* (Berlin: Walter de Gruyter, 2009): 404.

[9] Schmid, "Implied Reader".

[10] Crossover literature, in its most broad definition, includes texts that were written, published and marketed for a specific audience but widely read and adopted by readers of a different age.

[11] Beckett, *Crossover Fiction*, 181–87.

[12] Ibid., 196.

[13] Julia Eccleshare, *A Guide to the Harry Potter Novels* (London: Continuum, 2002): 7–8.

[14] Lindsey Fraser, *The Scotsman*, June 28, 1997.

[15] Philip W. Errington, *J.K. Rowling: A Bibliography 1997-2013* (London: Bloomsbury Academic, 2015): 46.

[16] "How We Rate and Review," Common Sense Media, July 28, 2020, https://www.commonsensemedia.org/about-us/our-mission/about-our-ratings.

[17] "Harry Potter Age-by-Age Guide," Common Sense Media, July 28, 2020, https://www.commonsensemedia.org/blog/harry-potter-age-by-age-guide.

[18] The value of the Lexile Measures is based on sentence length and word frequency, which are measures for syntactic and semantic difficulty respectively (Lennon and Burdick, 2004: 4).

[19] Derived from data on: https://hub.lexile.com/find-a-book/search and https://hub.lexile.com/lexile-grade-level-charts. The Lexile scores for the individual books in the Harry Potter serries are (chronologically): 880L, 940L, 880L, 880L, 950L, 920L and 880L.

[20] Errington, *J.K. Rowling*, 207.

[21] https://www.scholastic.com/teachers/teaching-tools/book-lists/the-complete-harry-potter-book-list.html#

[22] https://picarta.oclc.org/psi/xslt/DB=3.34/

[23] Stephen Krashen, "The Lexile Framework: Unnecessary and Potentially Harmful," *CSLA Journal* 24.2 (2001): 25–26.

[24] The analyses in this article are conducted on the digital edition Rowling, J.K. *Harry Potter: The Complete Collection* (2007), published by Bloomsbury. This edition contains corrections the publisher made to the texts in 2004. A complete overview of these corrections can be found here: https://www.hp-lexicon.org/differences-changes-text/

[25] Franz Wanner et al., "Are My Children Old Enough to Read These Books? Age Suitability Analysis," *Polibits* 43 (2011): 93–100.

[26] Alan Bailin and Ann Grafstein, *Readability: Text and Context* (London: Palgrave Mcmillan, 2016): 65-80.

[27] Colleen Lennon and Hal Burdick, "The Lexile Framework as an Approach for Reading Measurement and Success" (2004): 5.

[28] For this calculation, the English-dedicated sentence and word tokenizer of SpaCy (version 2.2.4, https://spacy.io) was used. In this regard, it is worth mentioning that direct speech of the type *"Yes," said Harry.* is considered as a single sentence, containing 3 words. For the purpose of tokenization at the word-level, all non-alphanumeric characters were discarded.

[29] The larger research project of which the current study is a part, tracks the construction of age in children's literature. One of the features that are studied is direct speech of fictional characters, attributed with the

corresponding age of the relevant character by manual annotation. The data pertaining to the ratio between direct and indirect speech in the *Harry Potter* series is derived from these annotations.

[30] Bailin and Grafstein, *Readability*, 13.

[31] Ibid, 65–80.

[32] The syntactic structure analysis was performed using the *Berkley Neural Parser* (Kitaev, Nikita, and Dan Klein, 'Constituency Parsing with a Self-Attentive Encoder.' In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 1, Melbourne* (2018)), which parses sentences using neural networks and self-attention. The model used in this analysis (benepar_en2) incorporates BERT word representations and achieves 95.17 F1 on the Penn Treebank.

[33] Victoria Johansson, "Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective," *Working Papers* 53 (2008): 61–79.

[34] Wendell Johnson, "Studies in Language Behavior: A Program of Research," *Psychological Monographs* 56.2 (1944): 1–15; Mildred C. Templin, *Certain Language Skills in Children; Their Development and Interrelationships* (Minnesota: University of Minnesota Press, 1957).

[35] Minimizing the effect of text length in order to gain insight into lexical richness has been the goal of many researchers within the field of vocabulary studies. Various alternative proposals have been made to replace TTR, or at least to accommodate for its drawbacks (Philip McCarthy and Scott Jarvis, "vocd: A Theoretical and Empirical Evaluation," *Language Testing* 24.4 (2007): 459-488. and Kristopher Kyle, "Measuring Lexical Richness," in Stuart Webb (Ed.) *The Routledge Handbook of Vocabulary Studies* (2019): 454-476. provide a non-exhaustive overview of such proposals). For the purpose of this article, it would lead us too far to investigate and apply all these alternative measures to the *Harry Potter* books.

[36] This means that we first calculate the TTR value for the segment of each book running from word 1 through 10000. Next, we calculate TTR for the subsequent window, which runs from word 2 through 10001, and so on. Once the TTR scores have been calculated for all segments that make up a book, we calculate the average score. As pointed out by Covington & McFall ("Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR)," *Journal of Quantitative Linguistics* 17.2 (2010): 94-100.), the calculation of MATTR varies with the chosen window size. For the purpose of determining the vocabulary size of an author, Covington and McFall recommend a window size as large as 10000 words (Covington & McFall, 2010: 97). The calculation of MATTR was performed, using 'lexical-diversity', a package for Python developed by Kristopher Kyle (https://github.com/kristopherkyle/lexical_diversity).

[37] Concretely, the ARI (1967) is calculated as follows: 4.71 * (number of characters / number of words) + 0.5 * (number of words / number of sentences) – 21.43.
Edgar A. Smith, Edgar and R.J. Senter, "Automated Readability Index," *Aerospace Medical Research Laboratories* (1967): 1–14.

[38] E. Fry, "Readability versus leveling," *The Reading Teacher* 56.3 (2002): 286-272.

[39] A. Schade, "The little read writing book: 20 powerful principles for structure, style, and readability," *Library Journal,* 129.13 (2004): 91.

[40] Scott A. Crossley et al., "Moving Beyond Classic Readability Formulas: New Methods and New Models," *Journal of Research in Reading* 42 (2019): 3.

[41] Gunning fog: Robert Gunning, *The Technique of Clear Writing* (McGraw-Hill, 1952);
Dale-Chall (original formula): Edgar Dale and Jeanne S. Chall, "A Formula for Predicting Readability: Instructions," *Educational Research Bulletin* (1948): 37–54;
Dale-Chall (revised version): Edgar Dale and Jeanne S. Chall, *Readability Revisited: The New Dale-Chall Readability Formula* (1995);

SMOG: G. Harry McLaughlin, “SMOG Grading: A New Readability Formula,” *Journal of Reading* 12.8 (1969): 639–46;
Coleman-Liau Index: Meri Coleman and T.L. Liau, “A Computer Readability Formula Designed for Machine Scoring,” *Journal of Applied Psychology* 60.2 (1975): 283–84;
Flesch-Kincaid formula: J. Peter Kincaid et al., “Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel,” *Research Branch Report 8-75* (1975).

[42] Reading speed is one of the components Crossley includes in his study of alternatives to classic readability measures (Crossley, “Moving Beyond Classic Readability Formulas”.)
See i.a. Bertram Bruce, A. Rubin & K. Starr, “Why readability formulas fail,” *IEEE Transactions on Professional Communication*, PC-24 (1981): 50-52;
Karen A. Schriver, “Readability formulas in the new millennium: what’s the use?,” *ACM Journal of Computer Documentation* 24.3 (2000): 138–140.

[43] The calculation of readability scores was performed using ‘readability’, a package for Python developed by Andreas van Cranenburgh (https://github.com/andreasvc/readability/).

[44] H. Mesmer, *Tools for matching readers to texts. Research-based practices* (Guilford Press, 2008): 26-27.

[45] T. M. Duffy, “Readability formulas: What's the use?,” in: T. M. Duffy and R. M. Waller (eds.), *Designing Usable Texts* (New York: Academic Press, 1985): 113-143.

[46] Mesmer, *Tools for matching readers to texts. Research-based practices*, 26.

[47] The code for this section of the article was developed by and in close collaboration with Mike Kestemont (University of Antwerp). Any mistakes or errors are our own. The code uses Python 3.6+ and has the following major dependencies: spaCy, numpy, sci-kit learn, pandas, lxml and SciPy.

[48] Yukun Zhu et al., "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books," *ArXiv* (2015).
This corpus, which is free to use for non-commercial research purposes, consists of self-published, novel-length fiction. The advantage of using this corpus to train a topic model with which to analyse the Harry Potter series is that it is contemporary, as opposed to the mostly historical work, free of copyright, that is often used in digital text analysis research. However, the self-published nature of these texts will result in stylistic differences to fiction that is mediated by publishers.

[49] Because this analysis is purely semantic, both the background corpus and the text of the series under investigation were pre-processed. The part-of-speech tagging feature of SpaCy was applied to the text to retain only the content words (nouns, verbs and adjectives). To further refine this schematic, a value of significance is given to each word. Terms that occur in fewer documents have a more specific semantics and thus receive a higher significance score. This weighted representation (TF-IDF model) is then used as data to build the topic model itself, using Non-Negative Matrix Factorization (NMF) as opposed to the equally popular Latent Dirichlet Allocation (LDA) method (see Mehdiyev et al. 2019). Both algorithms use the same input, but we favour the NMF method for this application as it is more stable in terms of parameter settings. We favour the NMF method for this application as it is more stable in terms of parameter settings. The model is trained by running through a fixed number of iterations in which it does two things. First the model calculates topic scores for each document based on the words it contains. These scores are then used to try and reconstruct the original words in the document. With each iteration, the model becomes more accurate.

[50] Dividing the series into chapters and putting these on a continuous scale presupposes that all instalments are equally distant from each other. We acknowledge that the variables used in this analysis might not be strictly continuous.

[51] Maria Nikolajeva, *The Rhetoric of Character in Children’s Literature* (Oxford: Scarecrow Press, 2002).

[52] Kate Behr, "'Same-as-Difference': Narrative Transformations and Intersecting Cultures in Harry Potter," *Journal of Narrative Theory* 35.1 (2005): 117.

[53] Jann Lacoss, "Of Magicals and Muggles: Reversals and Revulsions at Hogwarts," in Lana Whited (Ed.), *The Ivory Tower and Harry Potter: Perspectives in a Literary Phenomenon* (Columbia: University of Missouri Press, 2002): 67-88.

[54] Maria Nikolajeva, *Power, Voice and Subjectivity in Literature for Young Readers* (New York: Routledge, 2010).

[55] J.A. Appleyard, *Becoming a Reader: The Experience of Fiction from Childhood to Adulthood* (Cambridge: Cambridge University Press, 1994): 100.

[56] Behr, '"Same-as-Difference"', 114.

[57] Nikolajeva, *The Rhetoric of Character in Children's Literature*, 42.

[58] Appleyard, *Becoming a Reader*, 100.

[59] Beckett, *Crossover Fiction*, 67.

[60] Rebecca Butler, "The Literature Continuum: The Harry Potter Phenomenon," *School Libraries Worldwide* 9.1 (2003): 67.