

Research in Action: Constructing Age for Young Readers

VANESSA JOOSEN 

Children's literature studies has been relatively slow in adopting techniques from digital humanities. This article explains a method for digitising, annotating, and analysing texts in xml to investigate the implicit age norms that children's books convey. The case studies are seventeen books by Bart Moeyaert and La Belle Sauvage by Philip Pullman. The analysis of speech distribution, topic modelling, syntactic parsing, and lexical analysis with digital tools adds information about implicit age norms that can support and inspire narrative analyses with close reading.

Key words: *digital humanities, research methods, age studies, Bart Moeyaert, Philip Pullman*

In recent years, digital humanities has had a profound impact on literary studies, as demonstrated in a plethora of new publications, university programmes and research projects. The field of children's literature, however, has been rather slow to take part in this development, even though an increasing number of digitised children's books are being made available through virtual libraries (for example, the Baldwin collection in Florida or the Dutch digital library, DBNL). In this article, I introduce some of the techniques that we use in the research project Constructing Age for Young Readers, which runs at the University of Antwerp from 2019 to 2024 and is funded with a European Research Council (ERC) Starting Grant. This interdisciplinary research project combines theories and methods from digital humanities, children's literature studies, age studies, and empirical reader-response research to gain a better understanding of the ways in which age is constructed in children's literature on various levels. I focus specifically on how digital tools can support and inspire close reading to investigate the implicit age norms that the children's books in our corpus

International Research in Children's Literature 14.3 (2021): 252–268

Edinburgh University Press

DOI: 10.3366/ircl.2021.0409

© International Research Society for Children's Literature

www.eupublishing.com/ircl

convey. As case studies, I will first refer to a selection of seventeen books by the Flemish children's author Bart Moeyaert, winner of the Astrid Lindgren Memorial Award 2019, as well as *La Belle Sauvage* (2017) by Philip Pullman, a prequel to his famous *His Dark Materials* series. I demonstrate how topic modelling and lexical analysis with digital tools can supplement a narrative analysis of the implicit age norms in their works.

THEORETICAL FRAMING

The project Constructing Age for Young Readers (CAFYR) starts from the assumption that children's literature carries an ideological load (Hollindale; Stephens), including assumptions and ideas that are related to age. Researchers in age studies have pointed out that not only is age a biological given, but the meaning we give to age is also socially constructed (Green; Pickard; Gullette, *Aged by Culture*). Age is central to children's literature on various levels: the actors that are typically involved in the communication process of this discourse – adult authors and mediators, child and adolescent readers – are marked by an age difference. Moreover, children's literature is built around characters of all ages and often addresses age-related themes, such as growing up or intergenerational conflicts and collaboration (Deszcz-Tryhubczak and Jaques; Benner and Ullmann; Joosen, *Connecting Childhood*). Age critics and children's literature scholars have pointed out that some books for young readers transmit ageist prejudices and stereotypes, but children's books also have the potential to fight ageism and foster intergenerational understanding (Crawford; Henneberg; Deszcz-Tryhubczak and Jaques). Given the role that children's books play in the socialisation of the young, it is worthwhile to pay attention to the ideas that they convey with regard to age and to gain a better understanding of how these age norms are constructed and received. While it would be a mistake to read children's books as a mimetic rendition of reality, stories do contribute to the shaping of an age ideology. Stephen Katz even suggests that 'narrative is particularly important because it anchors the inside of aging, bringing together self and society and animating our biographies as we borrow, adapt, interpret, and reinvent the languages, symbols, and meanings around us to customize our personal stories' (n.p.).

Age norms are a central concept in the CAFYR project. Like gender, race, sexuality, and ability, societies develop norms to govern our understanding and performance of age. These can be made explicit and even take the form of laws (for example, the age of consent). More often, though, age norms remain implicit and need critical awareness to be noticed and debated. The CAFYR project aims for a better understanding of the age norms that children's books convey and questions what role age plays in the communication process from author to reader. Does an author's own age impact the way they think and write about age? Do authors who write for different age groups adapt the age norms they convey in a way that can be related to the age of the intended readership? Does the real reader's age affect the way they interpret age in children's books?

These are broad and challenging research questions that we try to answer using a combination of research methods. We interview readers of different generations who have read the same book about their views, interpretations, and reading experiences. We interview authors to ask about their ideas on their readership, as well as the impact of their own age on their writing as they have experienced it. We close read selected titles and use digital tools to capture bigger trends. Our corpus consists of 800 books: most are children's books, but the corpus also holds adult books by crossover authors who have written predominantly for children, such as J. K. Rowling, Anne Fine, and David Almond. These titles for adults allow us to draw comparisons with regard to the intended readership of the books. All the books were published after 1970. They appeared in the UK, the Netherlands, and Flanders. They are written by authors with a substantial oeuvre and a long writing career, who have published for readers in different age groups.

Part of the analysis of these narratives is carried out with the use of digital tools, which can be applied to large corpora and identify trends that would be more difficult or even impossible to detect with human research skills alone. To give a basic example: a computer script can yield the total number of words in a given book faster than most humans can count the letters in this sentence. Calculations are what computers are good at, and through calculations they are particularly well suited for pattern recognition, building models, and identifying salient features of (parts of) a corpus. In the CAFYR project, we both use third-party applications and develop our own computer scripts that help us identify stylistic features (stylometry) and distinctive topics (topic modelling), and engage in the exploration of affective states (sentiment analysis). In this article I will explain one method that we use to have computer scripts help us identify age norms in the construction of character speech. The tools help us map how the language of a certain fictional age group (e.g. children) is similar to or distinct from others (e.g. adolescents). Through the lexical features that characterise the speech of specific age groups we can also detect conversation topics that authors construct in relation to that age.

Due to the diversity of expertise needed and the efforts required to prepare the material, the nature of this research is highly collaborative. Our team includes researchers trained in programming in the computer programming language Python and in carrying out and interpreting computational, quantitative analyses.¹ For the digitisation and annotation process, we have relied on the help of volunteers. While the annotation of texts still involves human input and interpretation, we have set up guidelines and a feedback process to help the collaborators start from the same framework to limit highly subjective differences.

DIGITISING AND ANNOTATING CHILDREN'S BOOKS

Our digital analyses function on the basis of texts that are first digitised and then enriched with metadata and annotations. We received some novels in digital format through publishers, authors, and the Digital Library for Dutch Literature

(DBNL, www.dbnl.org). We scanned other novels ourselves and processed them with optical character recognition (OCR) software to create computer-readable text, which was then manually corrected. One limitation is that the images in the books are lost in this process. While software for digital analyses of images is currently being developed, it is not yet available in a form that is user-friendly enough for a project like ours, nor would it be able to capture the dynamic collaboration of texts and images that is so typical of illustrated books and picturebooks (Nikolajeva and Scott). We are keeping track of developments in this field, but for now, we exclude picturebooks from the digital analysis and only consider the illustrations in other books when we do a close reading of selected titles and discuss the books with real readers.

The plain-text files that result from the digitisation process are converted to xml files, a format that allows users to enrich the text with human-coded annotations. For this annotation process we work with two programmes: the freeware Sublime Text and the more sophisticated, but also more expensive, programme Oxygen XML Editor. The annotations help us to select specific parts of texts or even specific words and sentences for our analyses in later stages. We first add a header to the xml text for every book, containing information that is relevant to our project. Among this meta-information, we include the age of the author at the time of writing the book and the age of the intended reader as it is listed in Dutch library catalogues. These annotations are inserted in ‘tags’, standardised text between angle brackets that function as labels and help users extract specific elements of the text in a later stage. That way, we can tell our scripts to dedicate their attention to all books written by authors younger than twenty, for example, or to perform separate analyses for books for children and for adults, and compare the results.

We follow a standardised coding system developed by the TEI (Text Encoding Initiative, see Ide and Véronis), which has established encoding conventions to be used across projects. For Moeyaert’s *Blote Handen* [Bare hands], part of the header looks like this:

```
<listPerson type = ‘natural’>
  <person xml:id = ‘BM’>
    <persName>Bart Moeyaert</persName>
    <birth>1964</birth>
    <age when = ‘1995’ value = ‘31’>31 at the date of publication (1995)</age>
    <sex value = ‘M’>male</sex>
    <nationality key = ‘BE’>BE</nationality>
  </person>
</listPerson>
<listPerson type = ‘conceptual’>
  <person>
    <age>9</age>
  </person>
</listPerson>
```

For each instance of meta-information, we have an opening tag (e.g. <birth>) and a closing tag (e.g. </birth>) to respectively mark the beginning and ending of a labelled section. Several tags can be embedded on various levels, like Matryoshka dolls. We strive to use standardised tags (i.e. as they are recorded in the TEI guidelines) as much as possible. The tag <listPerson type='conceptual'>, for example, is used to record the age of the intended reader of the book. We have also customised some tags for our project's specific needs – for example, to calculate the author's age at the time of publication (<age when='1995' value='31'>).

In addition to these headers, we encode the text of the book itself to add information to selected words and passages, again with the aim of later extracting passages that fit certain criteria. We use tags to distinguish between direct and indirect speech and label all characters and references to characters (e.g. pronouns, nicknames) with a 'character ID'. A fragment of annotated text from Moeyaert's *Blote Handen* looks as follows:

```
<said direct='true' who='narrator'>On the other side, Elmer ran back and forth
sobbing. He's always detested swimming, and he was too small for jumping.</said>
<said direct='true' who='ward'>'Come on, Elmer!'<</said>
<said direct='true' who='narrator'><rs ref='ward'>I</rs>shouted.
Elmer cried harder than before and threw up his short front paws, and again
and again, but he didn't dare to jump. <rs ref='bernie'>Bernie</rs>and
<rs ref='ward'>I</rs>look at each other in the same moment. Big and
unwieldy like an ox, <rs ref='betjeman'>Betjeman</rs>moved closer. <rs
ref='betjeman'>He</rs>was already swearing.</said>2
```

Again, all marked text is contained between an opening tag <xxx> and a closing tag </xxx>, with various levels of embedding. In the tags starting with 'said' we include information about the speaker. Indirect speech where an extradiegetic third-person narrator is speaking would be labelled with the tag <said direct='false'>. *Blote Handen* has a first-person narrator, Ward. Since he is narrating the story retrospectively and may have aged since the moment when he experienced the events, we have created two character IDs for him: 'narrator' and 'ward'. The first tag: <said direct='true' who='narrator'> indicates that we have an instance of direct speech and that the narrator (Ward) is speaking. When we hear his voice at the time of the events, as he calls out for his dog Elmer, we indicate that this is direct speech expressed by the character in the story 'ward'. Further tags starting with 'rs' (short for 'reference string') encapsulate all human characters and pronouns in the xml file that refer to the characters – for example: <rs ref='ward'>I</rs>. We only tag references to animals if they are anthropomorphised and speak as humans would. This is not the case for Elmer.

The character IDs function as shortcuts to connect the annotated text with more elaborate information on the characters, which we collect in a spreadsheet. It would also be possible to enter all this information in the tag itself, but that would be too cumbersome. For each book, the annotator (a member of our research team or a student) makes a spreadsheet that matches the character ID (e.g. ward, bernie, betjeman in the quote above) with a set of identity features,

including age, gender, and ethnicity, and gives an indication of the character's role in the story (protagonist, friend, teacher, etc.).

Given CAFYR's aims, we are first and foremost interested in the category 'age', yet this information is often the most difficult of all character traits to trace. There are two reasons. First, while a character's gender and race usually remain stable throughout the story, it is quite common that fictional figures grow older in the course of a narrative, or that the plot contains flashbacks in which the characters are shown at a different age. In that case, we create several IDs for that character (e.g. wardbaby, ward8, ward12) and, if the novel is complex, we compose a timeline to keep track of the characters' ages at various points in the novel. We used this method for J. K. Rowling's *Harry Potter* series, for example, and for Philip Pullman's lengthy trilogy *His Dark Materials*. Second, characters' identity features are often not made explicit. Whereas we leave their gender and race unspecified when no information is given, in the case of age, we do try to assign all characters to an age category, and we try to be as specific as possible. We use exact numbers when a character's age is explicitly mentioned or can be easily derived. When that information is lacking, we refer to life stages that are rendered as refined as possible. Since various annotators are working alongside each other and we aim for maximum consistency, we have developed an age model that helps us to standardise the way the annotators attribute age to a character. The model has its roots in scholarly work from age studies, such as Lorraine Green's:

- unborn
- infant: ages 0 to 2
- child: ages 3 to 11
 - earlychild: ages 3 to 5
 - middlechild: ages 6 to 8
 - latechild: ages 9 to 11
- adolescent: ages 12 to 19
- adult: ages 20 and above
 - earlyadult: ages 20 to 39
 - twenties: ages 20 to 29
 - thirties: ages 30 to 39
 - middleadult: ages 40 to 59
 - forties: ages 40 to 49
 - fifties: ages 50 to 59
 - oldadult: ages 60 to 79
 - sixties: ages 60 to 69
 - seventies: ages 70 to 79

- deepoldadult: ages 80 to 100
 - eighties: ages 80 to 89
 - nineties: ages 90 to 100

Some level of interpretation is needed to categorise characters' ages according to this scheme. For example, if it is mentioned that a child is in the early years of primary school, the annotator can categorise it as a 'middlechild'. We ask annotators to indicate their level of certainty and give their reasoning behind assigning age categories. We recommend that if the level of certainty is low, they opt for a broad category. For *Blote Handen*, for example, Ward is assigned to the broad category 'child' with medium certainty – the annotator explains that he is called a child in the story. His mother is assigned to the overarching category 'adult' because she is a parent. Depending on Ward's age, she could be an old adult too, but this is less likely, and there are no indications in the text that would warrant this assumption. As this reasoning shows, the annotator's age norms inevitably come into play, which produces some circularity in our research: we investigate age norms but rely on our own age norms as we do so. By using broad categories, we try to minimise this effect.

ANALYSIS: DISTRIBUTION OF SPEECH

The annotated texts form the input for our digital analyses with Python scripts. Digital analysis can take various forms and cover widely different scopes. Matthew Jockers uses the term 'macroanalysis' for digital analyses on a large scale (for example all the digitised texts published in English in the nineteenth century), but digital analyses can also provide insights for a smaller corpus that might inspire new research questions and close readings of selected passages. Their capacities for counting, modelling, and comparing allow the computer scripts to pick up on aspects of the texts that readers will not consciously pay attention to or that would take scholars a lot of time to track (see, for example, Deijl et al.). Since our research is still a work in progress and annotations are time consuming, we have so far annotated and analysed sections of the corpus rather than the full body of texts. The results presented next are based on seventeen books by Moeyaert that we have annotated: sixteen children's books (in the broad sense of the term – that is, including his YA novels) and his novella for adults (*Graz*, 2009).

In the first stage of the analysis, we investigate the distribution of direct speech over different categories, which can give an indication of a character's importance and agency in a story. A script that we have written in Python easily calculates the percentage of direct speech uttered by every age group. The script first extracts all direct speech from our xml files (that is, all the text that we labelled with <said direct = 'true' who = 'character ID'> tags). It then connects the character IDs in the tags with the identity features in the spreadsheet. It

can thus link the extracted direct speech with the age that was attributed to the character. For the analysis presented here, we work with a simple binary model: child (for ages 0–19) and adult (for 20+). More specific subdivisions (e.g. middle child, age twelve, deep old adult) are automatically ranged under these broader categories. Next, all the extracted direct speech is grouped under the right age category, so that the script creates two files: one text file contains all the direct speech uttered by child and adolescent characters, and another holds all the direct speech of adult and old adult figures.

The distribution of direct speech across the characters' age groups looks completely different if we break down Moeyaert's oeuvre according to the age of the intended readership, as derived from library catalogue information. Figures 1 and 2 show these different distributions. Of course, you do not need a computer to establish that a novella like *Graz* features no child or adolescent speech and hardly any speech by older characters: this is self-evident if you read the actual book and pay attention to who speaks. It would take considerably more work to get an accurate picture of speech distribution in the sixteen books for young readers. From Figure 2, we learn that adults only account for a quarter of all speech. The difference between Figures 2 and 3 makes clear how heavily child narrators weigh in the distribution of speech. In Figure 2, first-person narration is included in the graph, while Figure 3 is based only on the dialogues in the text. While child and adolescent characters still utter the majority of words, the dialogues are more balanced. Moreover, Figures 2 and 3 also reveal that hardly any old adults speak in Moeyaert's children's books. This can raise the question of what role old age plays in his stories. After all, just because older characters rarely speak in his books, this does not mean that they cannot play significant roles – our analysis of direct speech does not capture how often older characters act in the stories or are on the minds of other characters. A comparison between Figures 2 and 3 also displays a remarkable trend in child and adolescent speech. If we include the narrator's speech in the distribution (Figure 2), we see that child speech dominates. If we only look at the dialogues (Figure 3), the adolescent characters take up a bigger portion of direct speech. This difference can inspire a closer investigation of the relationship between age and the narratological features of Moeyaert's work.

Since we have recorded not just age, but also other identity features, we can also measure the speech distribution of age and gender combined. Figure 4 adds the factor *gender* to Figure 3 (sixteen children's books by Moeyaert, narrator's speech excluded). From this graph, we learn that boys have more dialogue than girls in the category of child characters, but for the adolescents and adults, the picture changes. Adolescent and (old) adult speech is more dominated by female characters. Such findings can bring scholars to new insights on the level of an author's oeuvre. To me, they raise the awareness that, while male adolescents can still be found in Moeyaert's early YA novels, they seem to have disappeared from his later work, which features male child protagonists and female adolescent protagonists – something I had not realised before seeing

this quantitative analysis. It is also interesting to draw comparisons with other authors. Figure 5 displays the results for twenty-seven of Guus Kuijer's children's books: adults and old adults, especially male ones, take up a much bigger part of the dialogues. One explanation is the use of alter egos: Moeyaert is known to have based his first novels, when he was still an adolescent and early adult, on his own experiences (Lambrechts) and later often used his childhood as a source of inspiration; Kuijer, by contrast, has staged alter egos of his (old) adult self in various books (Joozen, *Adulthood*). The charts display striking differences that can also inspire further analyses of the authors' oeuvres and/or individual passages and books.

ANALYSIS OF LEXICAL VARIATION ACROSS AGE GROUPS

To continue the report of our digital analyses, we want to know not only if there is a distinction in the distribution of speech over different age groups, but also whether the words uttered vary and thus reveal insights into the construction of age. From the file with the speech of children/adolescents and the file of (old) adults' speech used for the distribution calculation, we first remove all character names automatically (on the basis of the names of characters recorded in the spreadsheet), as well as highly frequent stop words that mainly serve grammatical functions (such as articles and the most common pronouns). After a test run, and based on the discrepancy between Figures 2 and 3, we decided to remove the direct speech of the narrator because this weighed heavily on the results and created a distorted picture, with a high amount of words to introduce direct speech and describe basic actions in the story (for example, with words like 'went', 'get'). For similar reasons, we excluded the only novella for adults, *Graz*, because it dominated adult speech with words that could only be found in that book (e.g. *mayfreddygasse*, a street in Graz; *portefeuille* [wallet]; *apotheker* [pharmacist]; *sneeuw* [snow]; *poort* [gate]). Instead, we decided to analyse only the words uttered in dialogues in Moeyaert's children's books.

We then implemented the open-source tool Scattertext, developed by Jason Kessler, that helps to visualise how two sets of text differ from each other (Figure 6). The X-axis charts the frequency of words in adult and old adult speech, while the ranking on the Y-axis reflects the occurrence of words in child and adolescent speech. This means that the higher up a particular dot is on the vertical axis, the more frequently it appears in child or adolescent speech. Similarly, the further right on the horizontal axis, the more a dot's corresponding word is used by an (old) adult character. Words in the top left corner are often used by children, but rarely by adults; those in the bottom right corner are frequent in adult speech, but not in children's speech; and in the top right corner we see the words that are common to both groups. When clicking on the dots, the chart allows for an interactive exploration of the context in which the words are used and you can see which character utters them.

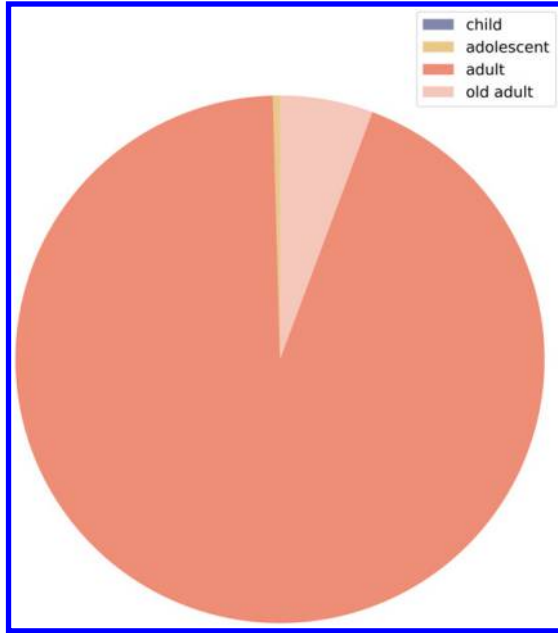


Fig. 1. Distribution of speech in *Graz*.

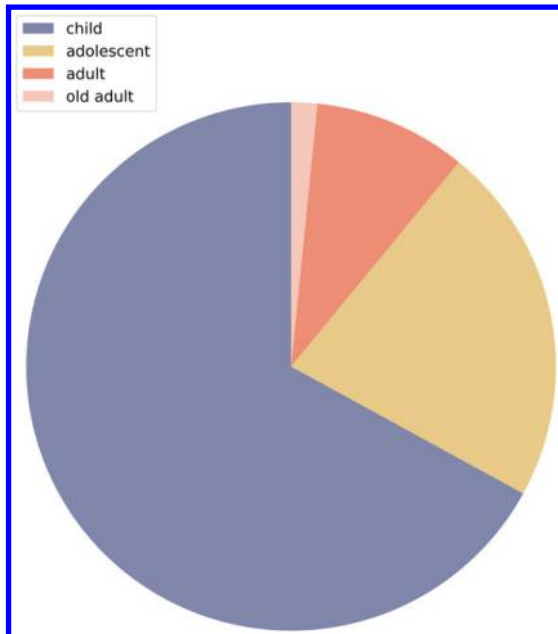


Fig. 2. Distribution of speech in Moeyaert's children's books with narrator's speech included.

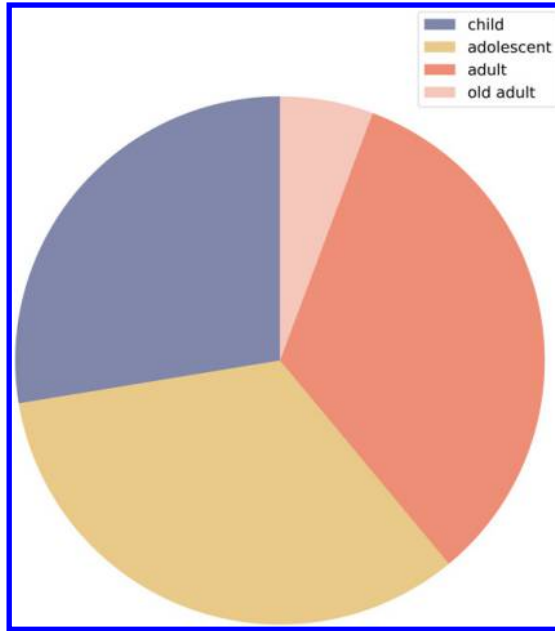


Fig. 3. Distribution of speech in dialogues in Moeyaert's children's books.

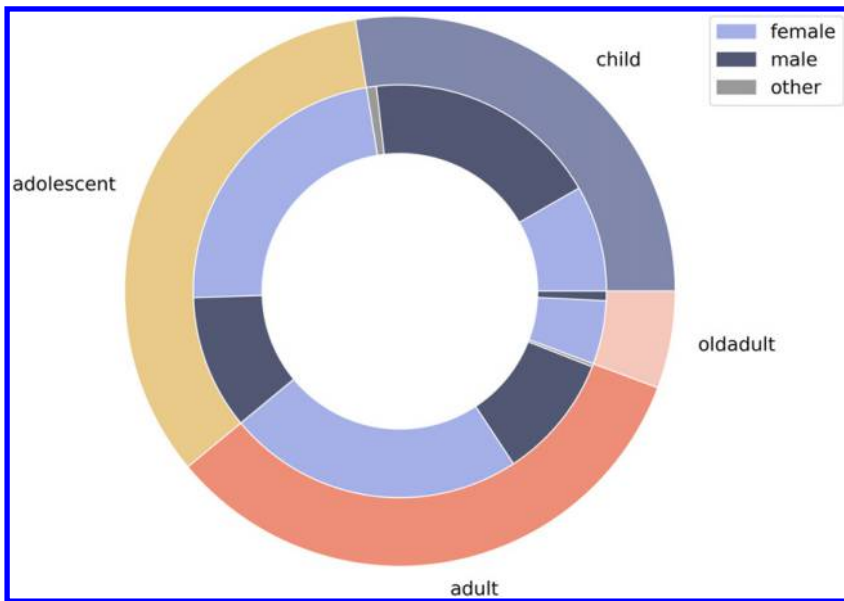


Fig. 4. Distribution of speech according to gender and age in Moeyaert's children's books.

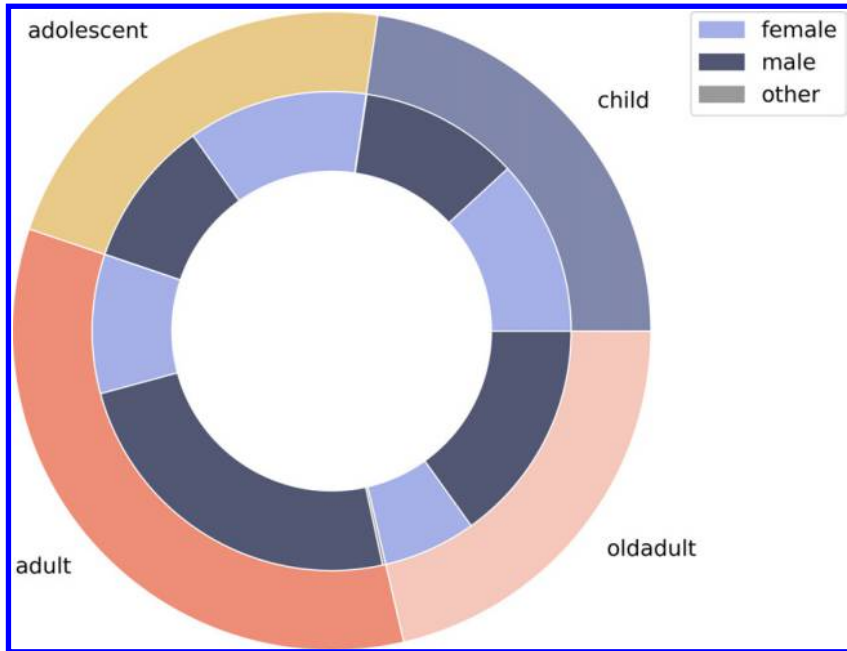


Fig. 5. Distribution of speech according to gender and age in Kuijer’s children’s books.

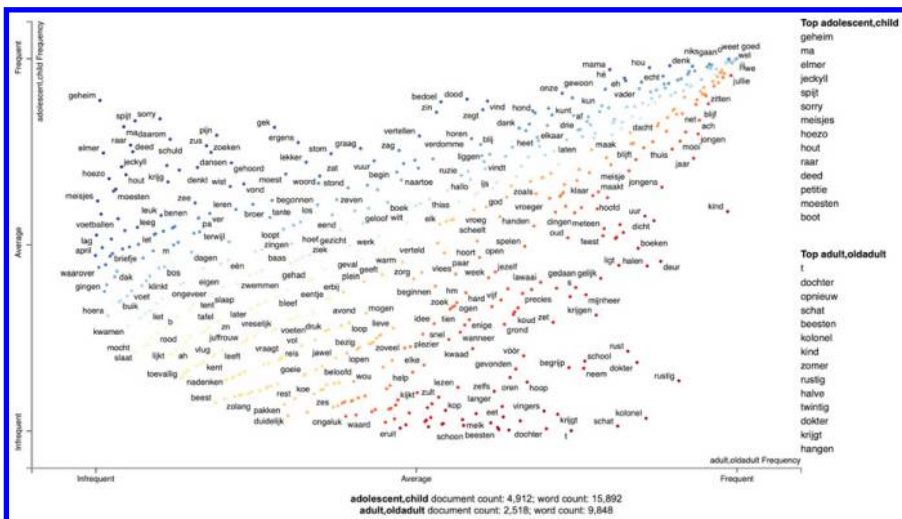


Fig. 6. Scatterplot with direct speech in Moeyaert’s books for young readers.

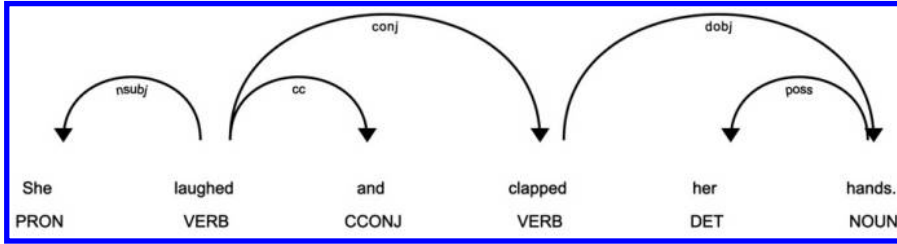


Fig. 7. Syntactic dependency parse tree of an example sentence from *La Belle Sauvage*.

The most common distinctive words for children and adolescents are: *geheim* [secret], *smacht* [longing], *ma* [mum], Elmer [name of dog], *briefje* [little note], *sorry* [sorry], *spijt* [regret], *raar* [strange], *hout* [wood], *petitie* [petition], *deed* [did], *hoezo* [how come, what?], *moesten* [had to], *boot* [boat], *zus* [sister], *pijn* [pain], *voel* [feel], *denkt* [thinks], *wist* [knew], *dansen* [dancing], *zoeken* [searching]. For (old) adults, the most common idiosyncratic words are: *'t* [short form of it], *dochter* [daughter], *opnieuw* [again], *schat* [dear, darling, treasure], *kind* [child], *beesten* [beasts, animals], *colonel* [colonel], *rustig* [calm], *winter* [winter], *zomer* [summer], *halve* [half], *twintig* [twenty], *begrijp* [understand], and *krijgen* [get].

At this point, human interpretation is needed to formulate hypotheses and to integrate the scatterplot visualisation with other insights from research and/or close reading. The words *geheim* [secret] and *smacht* [longing] are used almost exclusively in single titles, *Kus me* and *Echt weg is niet zo ver* respectively. By contrast, the words related to apologies and regret clearly surface in the child and adolescent speech, and if we click on the dots, we see that they are not unique to a single book but appear throughout Moeyaert's oeuvre. The twenty-eight mentions are distributed over fourteen different young characters; only two adult characters apologise in this way (with one mention for each). Similarly, the word '*spijt*' [regret] is used twenty-seven times by sixteen different young characters versus only two adult ones. It often features in the expression '*het spijt me*' [I am sorry]. This indicates that child characters apologise or express regret more than adult ones—a hypothesis that would be interesting to test in close reading, where it can be more contextualised. Kessler's scatterplot comes with a search box that helps you locate specific words on the scatterplot. A search for the word '*schuld*' [fault, guilt] shows that it is also situated close to the top left corner: it is articulated eighteen times by various young characters and only three times by (old) adult ones. The provided context already nuances the previous findings because it is often part of the expressions '*jouw schuld*' and '*jullie schuld*' ([your fault] singular and plural). This suggests that the young characters do not just apologise, but also defend themselves against accusations.

From the most common words listed here, you can also derive that intergenerational communication is a recurrent feature in Moeyaert's dialogue, with children talking about or addressing adults (mum) and vice versa (child, daughter in adult speech). A specific search shows that '*ma*' [mum] is addressed or mentioned far more by child figures (twenty-six mentions) than its male

counterpart ‘*pa*’ ([dad] twelve mentions). Although this distinction is less evident in the related words ‘*mama*’ (seventy-three) and ‘*papa*’ (sixty-three), it matches with the impression above that female adult figures take up a bigger part in the dialogues than male adult figures. Two possible hypotheses can result from this that may guide a closer analysis of the books: that women, and mothers in particular, play a more prominent role in Moeyaert’s fiction than fathers, or that the latter are more silent. Keeping the results of speech distribution (Figure 3) in mind, it is also striking to note that the terms ‘*oma*’ and ‘*opa*’ do not feature as a distinct feature of child speech in Moeyaert’s scatterplot. This confirms the impression that older characters do not play a significant role in his books. The word ‘*oude*’ (old) does appear, but it is mainly used by adults and rarely refers to people.

ANALYSIS WITH PARSER

In addition to this analysis of speech distribution and frequently used words, we have developed an automatic syntactic dependency parser, which takes the annotated xml file as input in combination with the spreadsheet containing character information. Automatic syntactic parsing, where the script automatically extracts words from the text that fulfil certain grammatical functions, has already proven its use in computer-aided literary research. For example, through automatic parsing Bamman et al. investigated the stereotypical actions and attributes associated with movie characters in over 42,000 movie plot summaries. And for a corpus of thirty-two novels, Koolen and Van Cranenburgh automatically extracted descriptions of characters’ physical appearance, establishing a difference between such descriptions in chick lit and other literary novels. CAFYR draws inspiration from this research and investigates age norms through the adjectives, verbs, and possessives that are associated with characters belonging to certain age groups. We can already use this parser for English; the Dutch counterpart is under construction. Consider, for example, the following sentences from Philip Pullman’s *La Belle Sauvage*:

```
<rs ref='fenella'>She</rs>laughed and clapped <rs ref='fenella'>her</rs>
hands. In the pale sunlight that came through the dusty windows, <rs
ref='malcolm'>Malcolm</rs>saw how chapped and cracked the skin of
<rs ref='fenella'>her</rs>fingers was, how red and raw. <said direct='true'
who='adnan'>'Lot of people in tonight?'</said><said direct='false'>said the
dark-eyed <rs ref='adnan'>man</rs>.</said>
```

All the sentences are syntactically ‘parsed’ with the open-source software SpaCy (Honnibal and Johnson). This software can recognise grammatical dependencies. For the first sentence in the fragment above, this renders the picture shown in Figure 7. When matched with our annotations and the character traits in the spreadsheet (linked via the character ID), the parser can extract relevant information for specific characters and for age groups. The script will capture, for example, that ‘laughed’ is an action performed by Fenella (because ‘She’

has been tagged as such in the xml) and that Fenella is an old adult (based on information from the spreadsheet with character traits). Likewise, 'hands' is counted as one of her possessions. It records that Adnan (an adult) is dark eyed and called 'a man', and that Malcolm (age eleven according to the character list) performs the action of seeing.

If we extend this analysis to the full text of *La Belle Sauvage*, we find that the following adjectives predominate in relation to child characters: 'little', 'young', 'sleeping', and 'curious' – perhaps not surprising given the presence of baby Lyra. Top adjectives modifying adolescents are: 'little', 'sleeping', 'scrawny', 'sodden', 'solitary', and 'senior'. Adult characters are most often accompanied by the adjectives 'other', 'young', 'dead', and 'gyptian' (a particular ethnic group in the novel), and old adults by 'old', 'gyptian', 'elderly', 'good', and 'poor'. Some of these adjectives are clearly book specific ('gyptian' and 'sodden', as the story takes place during a flood). Other common age norms are also visible here: child curiosity, adolescent solitude, and the decline narrative of old age (Gullette, *Age-wise*). A larger corpus is needed to see if trends emerge that are specific to authors, age groups, time periods, and/or intended readership. More data also allow further refinement according to gender and ethnicity, or the character's role in the story.

CONCLUSION


In the first phase of the project, we have focused on developing scripts and methods, and carried out analyses on subparts of the corpus. The next stage involves the application of these tools to larger sets (including the full corpus of 800 books) and to draw comparisons on the basis of various parameters. A big challenge will lie in making a wise selection from an overabundance of possible combinations and refinements. Our overarching research questions, in particular concerning the impact of the age of the author and the intended reader, will guide this selection. In addition, a selection of the most salient trends and hypotheses will guide a close reading of selected passages and titles. At that point, we can also supplement the numerical, categorical approach to age of the digital tools with other perspectives.

The analyses that I have presented make use of advanced techniques and require the input of experienced colleagues in digital humanities. For analyses on a smaller and more basic scale, user-friendly freeware is available online. Laurence Anthony's Antconc and Voyage tools (<https://voyant-tools.org/>) in particular allow users to compare the vocabulary of two (sets of) texts, for example, or give quick overviews of words in context. The results of such tools, as well as the scripts that we have applied, cannot replace a more literary and contextualised reading of characters and passages, but they offer starting points for considering those texts with fresh eyes and getting a fuller understanding of them.

ACKNOWLEDGEMENTS

The author wrote this article as part of the research project Constructing Age for Young Readers. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 804920). The author would like to thank Mike Kestemont, Wouter Haverals, and Lindsey Geybels for their support in developing the research that forms the basis of this article, as well as the students who helped with the annotations of the primary texts.

ORCID

Vanessa Joosen  <https://orcid.org/0000-0001-8060-5728>

NOTES

1. The code that we have developed for the project is available through open access via Zenodo, together with a detailed manual for the annotation process. The DOI for the code is: 10.5281/zenodo.5105898. While we cannot share the digital files of primary works due to copyright, the code is developed in such a way that the quantitative analyses can also be conducted with other digital text corpora. This way, we aim to cater to the interests of other scholars who want to pursue quantitative analyses in their own research.
2. My translation. Original text: 'Elmer liep aan de overkant jankend heen en weer. Aan zwemmen heeft hij altijd een hekel gehad, en voor springen was hij te klein. "Kom dan, Elmer!" riep ik. Elmer huilde harder dan eerst en gooide zijn korte voorpootjes in de lucht, en nog eens en nog eens, maar springen durfde hij niet. Bernie en ik keken elkaar op hetzelfde moment aan. Groot en log als een os kwam Betjeman dichterbij. Hij vloekte al.'

WORKS CITED

- Bamman, David et al. 'Learning Latent Personas of Film Characters'. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (2013): 352–61.
- Benner, Julia and Anika Ullmann. 'Doing Age und die Relevanz der Age Studies für die Kinder- und Jugendliteraturforschung'. *Fakt, Fake und Fiktion: Jahrbuch der Gesellschaft für Kinder- und Jugendliteraturforschung*. Eds Gabriele von Glasenapp, Emer O'Sullivan, Caroline Roeder, Michael Staiger, and Ingrid Tomkowiak. Berlin: GKJF, 2019. 145–59.
- Crawford, Patricia. 'Crossing Boundaries: Addressing Ageism through Children's Books'. *Horizons* 40.3 (2000): 161–74.
- Deijl, Lucas, Saskia Pieterse, Marion Prinse, and Roel Smeets. 'Mapping the Demographic Landscape of Characters in Recent Dutch Prose: A Quantitative Approach to Literary Representation'. *Journal of Dutch Literature* 7.1 (2016): 20–42.
- Deszcz-Tryhubczak, Justyna and Zoe Jaques (eds). *Intergenerational Solidarity in Children's Literature and Film*. Jackson, MS: University of Mississippi Press, 2021.
- Green, Lorraine. *Understanding the Life Course: Sociological and Psychological Perspectives*. Cambridge: Polity Press, 2010.
- Gullette, Margaret Morganroth. *Aged by Culture*. Chicago: University of Chicago Press, 2004.
- . *Age-wise: Fighting the New Ageism in America*. Chicago: University of Chicago Press, 2011.
- Henneberg, Sylvia. 'Of Creative Cronos and Poetry: Developing Age Studies through Literature'. *NWSA Journal* 18.1 (2006): 106–25.
- Hollindale, Peter. 'Ideology and the Children's Book'. *Signal* 55 (1988): 3–22.

- Honnibal, Matthew and Mark Johnson. 'An Improved Non-Monotonic Transition System for Dependency Parsing'. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (2015): 1373–8.
- Ide, Nancy and Jean Véronis (eds). *Text Encoding Initiative: Background and Contexts*. Dordrecht: Springer, 1995.
- Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press, 2013.
- Joosen, Vanessa. *Adulthood in Children's Literature*. London: Bloomsbury, 2018.
- . (ed.). *Connecting Childhood and Old Age in Popular Media*. Jackson, MS: University of Mississippi Press, 2018.
- Katz, Stephen. 'What Is Age Studies?' *Age Culture Humanities* 1 (2014). 28 August 2020. <<http://ageculturehumanities.org/WP/what-is-age-studies/>>.
- Kessler, Jason S. 'Scattertext: A Browser-Based Tool for Visualizing How Corpora Differ'. 2017. 1 September 2020. <<http://arxiv.org/abs/1703.00565>>.
- Koolen, Corina and Andreas van Cranenburgh. 'Blue Eyes and Porcelain Cheeks: Computational Extraction of Physical Descriptions from Dutch Chick Lit and Literary Novels'. *Digital Scholarship in the Humanities* 33.1 (2018): 59–71.
- Lambrechts, Johan. 'Sprookjes zijn ook leugens'. *Top*, 30 January 1987: n.p.
- Moeyaert, Bart. *Blote handen* [Bare hands]. Amsterdam: Querido, 1995.
- . *Broere* [Brothers]. Amsterdam: Querido, 2002.
- . *Dani Benmoni*. Amsterdam: Querido, 2004.
- . *De melkweg* [The milky way]. Amsterdam: Querido, 2011.
- . *Duet met valse noten* [Duet with false notes]. Averbode: Altiora, 1983.
- . *Echt weg is niet zo ver* [Far away is not that far]. Tilburg: Zwijzen, 1993.
- . *Een klap is geen kus* [A slap is no kiss]. Tilburg: Zwijzen, 1989.
- . *Graz*. Amsterdam: Querido, 2009.
- . *Het boek van Niete* [Niete's book]. Averbode: Altiora, 1988.
- . *Het is de liefde die we niet begrijpen* [It's love we don't understand]. Amsterdam: Querido, 1999.
- . *Iemands lief* [Someone's love]. Amsterdam: Querido, 2013.
- . *Kus me* [Kiss me]. Averbode: Altiora, 1991.
- . *Mansoor*. Amsterdam: Querido, 1996.
- . *Missen is moeilijk* [Missing is hard]. Amsterdam: Querido, 2008.
- . *Suzanne Dantine*. Averbode: Altiora, 1989.
- . *Tegenwoordig heet iedereen Sorry* [These days everyone is called Sorry]. Amsterdam: Querido, 2018.
- . *Terug naar af* [Back to square one]. Averbode: Altiora, 1986.
- . *Wespennest* [Hornets' nest]. Amsterdam: Querido, 1997.
- Nikolajeva, Maria and Carole Scott. *How Picturebooks Work*. New York: Garland, 2001.
- Pickard, Susan. *Age Studies: A Sociological Examination of How We Age and Are Aged Through the Life Course*. Los Angeles: Sage, 2016.
- Pullman, Philip. *La Belle Sauvage*. Oxford: David Fickling, 2017.
- Stephens, John. *Language and Ideology in Children's Fiction*. London: Longman, 1992.

Vanessa Joosen is associate professor of English literature and children's literature at the University of Antwerp. There she leads the ERC-funded project Constructing Age for Young Readers and organises the annual Children's Literature Summer School. Vanessa Joosen is the author of, amongst others, *Adulthood in Children's Literature* (Bloomsbury, 2018) and edited the volume *Connecting Childhood and Old Age in Popular Media* (University of Mississippi Press, 2018).