University of Antwerp

Faculty of Arts
Department of Literature

Reading the Reader from a Distance

Digital Analyses of the Construction of Age in Fiction for Different Ages

Thesis submitted for the degree of
Doctor in Literary Studies
at the University of Antwerp

To be defended by

Lindsey Geybels

Supervisors
Prof. dr. Vanessa Joosen
Prof. dr. Mike Kestemont

Antwerp, 2023

Table of contents

Chapter 1 General Introduction

Children's literature has long been acknowledged as a powerful vehicle for shaping young minds and conveying ideological messages (Hollindale *Ideology*, and Stephens *Language*). Scholars in the field of children's literature studies have extensively explored the ways in which literature for children carries this ideological load, including the way in which age is constructed and represented. Age ideologies embedded in fiction for young readers play a significant role in the age socialisation process, intertwining with real-life experiences to shape children's understanding and perception of age within a broader social context (Gullette, *Aged by Culture*). These ideologies are incorporated in fiction in both implicit and explicit ways, which can aid children in learning to interact with others but also potentially perpetuate harmful biases. Ageism, traditionally defined as the presence of negative stereotypes about old age (Ansello), extends beyond prejudice against older people and encompasses discrimination against individuals of any age. As ageism has been shown to manifest already at a young age (see Bodner, and Montepare and Zebrowitz), a critical examination of children's literature can provide valuable insights into the age ideologies at play. Moreover, age ideologies in fiction impact individuals of different age groups, including children, young adults, and adults. Acknowledging the different age ranges that children's literature is aimed at, from babies to young adults, is important to gain a deeper understanding of the influence the age of the intended reader has on the construction of age in fiction.

The concept of age plays a diverse role in children's literature, even more so than in adult literature, and is an important component both of definitions of the genre and of the discussion of its readership. Children's literature researchers have offered different perspectives on what constitutes the material they study, with some emphasising its target audience and defining it as literature explicitly created for children (Nodelman) or read by children (Lerer). This view on the definition of children's literature is complicated by the existence of distinct terms used to describe different types of readership, including the real reader, the implied reader, the model reader, and the intended reader. In this dissertation, the focus lies on the concept of the intended reader — the reader that an adult author or mediator, be it the editor, publisher, or other gatekeeper, have specifically targeted to engage and influence through the literary work. Rather than adhering to a strict definition of children's literature, other scholars emphasise the traditional characteristics inherent to this genre. Fiction for young readers often includes characters representing a wide spectrum of ages, and frequently explores themes that are closely tied to age, such as the transformative journey of maturation (which it inherits from the Bildungsroman) or the dynamics involved in intergenerational relationships (Joosen, 'Second Childhoods'). This study takes a combined approach to ideology; in addition to considering the sociological component of the age of the intended reader, it examines the literary representations of age in fiction from a discursive standpoint, focussing on the actual text and building on the assumption that "ideology is inherent to every form of discourse and every utterance" (Herman and Vervaeck 'Ideology', n.p.).

To delve into this multifaceted construction of age in fiction, this dissertation employs tools from the field of digital humanities. The examination of age ideologies in children's literature is not new. However, it has traditionally involved studying children's literature as a cultural artifact,

thereby introducing an inherent subjectivity that accompanies the researcher's interpretation as well as a limitation on the scope of the material that can be analysed. Through the application of computational tools, I approach the study of the construction of age in fiction for young readers from a fresh perspective by transforming literature into analysable data. This approach does not seek to supplant one method with another, but rather aims to complement existing approaches by offering novel insights and enhancing our understanding of age ideologies within children's literature. By examining a large set of texts targeted at children of varying ages as well as texts for adults, the aim of this research is twofold.

RO 1 Gain deeper understanding of the construction of age in literature for young readers.

RO 2 Compile a (non-exhaustive) set of methods for digital text analysis that enables the study of age-related questions, both on the level of the intended reader and of the explicit and implicit ideology expressed in fiction.

These research objectives contribute to the project Constructing Age for Young Readers (CAFYR), which is led by prof. dr. Vanessa Joosen at the University of Antwerp and funded with a European Research Council (ERC) Starting Grant. The project aims to enhance our comprehension of the age norms embedded within children's books and to explore the significance of the age of both authors and readers in a corpus of Dutch, Flemish, and British fiction for young readers published between 1970 and 2020. It investigates whether an author's own age influences their perspective on and writing about age, if the real reader's age influences their interpretation of age portrayed in children's books, and whether authors (in collaboration with other mediators) adapt the age norms they convey in their work based on the intended readership's age. The remainder of this introduction contextualises the latter objective within a theoretical framework and elaborates on the methodology employed in this thesis, following the process from building a corpus and text annotation to an overview of the analyses used in subsequent chapters.

1.1 Theoretical framework

Before addressing the objectives and specific research questions that form the guiding principles of this dissertation, the following section discusses the context and theoretical frameworks from which they emerge. The key concepts featured throughout this dissertation and underpinning all analyses are those of intended readership, the construction of age, and the intersection between children's literature studies and digital humanities.

1.1.1 Intended readership

As pointed out in the previous section, the concept of readership is a complex one. Scholars employ varying terminologies to describe the concept of the reader, distinct from the actual reader, as they perceive the abstract reader's role differently, such as the ideal reader, the superreader, the model reader, or the implied reader. While some approaches focus on the reader as an interpretive device, others emphasise the reader as the intended recipient of the text. One example falling into the former category is Umberto Eco's model reader, which Eco deems an essential component of any text. This model reader aligns with a set of conditions that must be met for the text to be comprehensible (11). The text itself emerges through the integration of information provided by the author and that which is contributed by the model reader, with the latter being influenced by the author's input (206). In contrast, Wolfgang Iser's concept of the implied reader aligns with the notion of the reader as an addressee and is more closely related to the reader's role as discussed in this dissertation. Iser's implied reader can be likened to Wayne Booth's implied author, both operating within and beyond the text as mechanisms for constructing meaning (Prince, n.p.). However, Iser's contention that the implied reader fills intentional gaps in the text has drawn criticism for overemphasising the role of textual input and neglecting the inherent interpretative freedom in the act of reading (Mailloux, 51–53). This has led to the perception of the implied reader as equivalent to authorial intention or textual meaning (Prince, n.p.).

The term adopted throughout this thesis, with the exception of chapter three (for reasons that I will explain below), stems from Perry Nodelman's definition of children's literature on the one hand and Rita Ghesquière's discussion of the term on the other, albeit with some reservations and modifications. Nodelman notes that the categorisation of texts as children's literature is done on the basis not of the authors of those texts, which is the case with most other types of literature, for example African American literature, but rather on the criterion of their intended readership (3). Children's literature is not commonly produced by children, but written and published for younger readers, a term I use here to describe fiction aimed at young people of all ages, from infant to adolescent. According to Ghesquière, this concept of the young reader is compiled from an author's memories of their own childhood, their day-to-day encounters with children and their ideologies (88) and it guides the author in their writing process. The main characteristic of the intended reader that emerges from most discussions of intended readership, and the one on which this thesis focusses as well, is that of age. While subgenres of children's literature can also be based on gender (girls' literature and boys' literature), this categorisation, as well as a possible classification based on, for example, ethnicity or class, is rarely addressed by authors or explicitly printed on books, a practice that is more common with age markers.

The emphasis that Ghesquière's definition places on authorial intent can also be found in Peter Hunt's conception of the age of the intended reader as informing the decisions authors of children's literature make in terms of language and context (548). This way of looking at the intended reader is reflected in the poetics of several of the authors included in the corpus of this thesis. Anne Fine, who publishes books for readers of different ages in the United Kingdom, for example, has a specific age in mind from the moment she starts conceptualising each book that she writes, even if that age marker often does not appear explicitly on the published book (Fine, *Interview with Anne Fine*). Anna Woltz, an author included in the Dutch part of the corpus, has a similar writing process. She writes her children's books on the premise that child readers only enjoy stories in which the main characters are the same age or slightly older than themselves (Nikolajeva *Rhetoric*, 271). For her, the age she chooses for her protagonist simultaneously determines the age of the intended reader (Woltz, *Interview with Anna Woltz*). However, authors do not always get to decide the age for which their books are published. Aline Sax, another Dutch-language author in the corpus, also typically writes with an intended age in mind, but when she sent a book she initially wrote for adults with reading difficulties to a publisher of children's literature, it was decided, without her involvement, that this book was suitable for children aged ten and above (Sax, *Interview with Aline Sax*). Moreover, the German translation of this book, *Het meisje en de soldaat* (2013), was published as an age-less book, without any mention of an intended age.

Authors are thus not the only mediators in the process of producing books for young readers who decide on the appropriate age of the intended readership of their texts. Indeed, there are many authors who do not have a specific age in mind while writing, either because they rather leave it up to their publisher (e.g., Bart Moeyaert, *Interview with Bart Moeyaert*), they write only for themselves (e.g., Philip Pullman, 'Questions'), or they see the age of the intended reader as something that is to be discovered during the writing process (e.g., David Almond, *Interview with David Almond*). The common conviction of writers who do not address the age of their intended readership is that writing with a specific age in mind puts limits on the artistic process and, as a result, will always get in the way of a good story (Hunt). In these cases, publishers, librarians, and bookshops determine which category of reader they want to aim the book at. This practice has existed since the beginning of printing in the 15th century when William Caxton added guidelines for the intended reader, including for "yonge children", to the books he printed (qtd. in Rassuli). For a long time, it was also common to include other characteristics of the intended reader, such as gender and class, in the title of a prose text (Ghequière, 89). Whoever is responsible for this categorisation aside, stipulating the age of the intended reader of a book already reveals that the society in which the text is written and published has certain expectations about readers of different ages, about the difficulty of the texts they can read, and what topics they are interested in.

The way the term 'intended reader' is used in this thesis facilitates the study of these ideologies in books for children of different ages as well as for adults. First, like Nodelman, I consider it the defining factor in the classification of literature based on age, but instead of adhering to the dichotomy between children's and adult literature, I consider the further segmentation within children's literature, ranging for books for children aged 7 to young adult literature. Although literature specifically designed for children under the age of 7, such as baby books or first reader books, exists, I chose not to include these in the corpus of this dissertation because such publi-cations differ significantly in style and contain minimal textual content. Consequently, conducting

reliable digital analyses on such materials presents challenges due to the limited amount of data that can be extracted from them (also see section 1.2.1). Since a similar age-based classification does not exist within adult literature, all works written for this age group are studied as belonging to the same category. Second, I do not consider the determination of these age markers to be entirely in the hands of the author. While author websites were consulted during the phase of corpus building, I also refer to publisher's websites and library databases to collect metadata. The only chapter that does not use the concept of the intended reader as delineated above is chapter three, which is based on an article I co-authored with dr. Wouter Haverals, a former team member of the *Constructing Age For Young Readers* project. Following the advice of the peer reviewers of this publication, we revised our use of the term 'intended reader' to 'implied reader'. Indeed, the focus of the research reported in the article is the reader as constructed by the text in an effort to study the readership of a series of books whose author does not declare their intended age. Informed by one of Wolf Schmid's views of the concept, the implied reader is here understood as the ideal reader, the person who understands the text. However, most of the chapters in this thesis rather follow another conception proposed by Schmid, which lies closer to the addressee of the text, whose "linguistic codes, ideological norms, and aesthetic ideas must be taken into account" (n.p.). This interpretation of the intended reader does not stem from the text itself, but is determined by an outside entity, namely the author, publisher, or other literary institution.

Authors often adapt their writing style unconsciously, both those who are aware of the age they write for and those who say they do not write with a specific readership in mind. Additionally, authors who target their writing to readers of a specific age consciously adapt their work in several ways. Recalling Fine's poetics above, she mainly considers the content of her work in relation to the age of the intended reader. For her, it is important that she writes on topics that will captivate her readers; her adult books therefore usually contain "topics that are of less interest to children" (*Interview with Anne Fine*). She also tailors the length of the work, writing shorter stories for younger children than for slightly older children. Other authors primarily adapt their vocabulary to the age of their intended readership. For some, writing for younger children feels limiting because of this and, in the case of Jacqueline Wilson, writing for older children makes her "feel free to develop in all kinds of directions" (*Interview with Jacqueline Wilson*). Not only do authors themselves observe these differences between their works addressed to readers of different ages, literary scholars, mainly within the field of children's literature studies, have often scrutinised such texts to gain a greater understanding of the nature of children's and adult literature. By comparing full stories or passages from books for children on the one hand and adult novels on the other, researchers including Anderson, Appleyard, Beckett (*Crossover Fiction*), Ghesquière, Grit, Hurkmans, Joosen ('Counting Stars'), Shavit (*Poetics*), Stephens ('Maintaining'), Talley, and Van Lierop-Debrauwer ('Grote Gelijkenis'), reached insights related to differences in, among others, writing style, genre, content, and layout. This dissertation contributes to this research in two innovative areas. First, it does not consider children's literature as a single category, but rather takes into account further segmentation, as described above. Second, it builds on the methodology employed by a small number of previous studies; instead of studying a small corpus through the established method of close reading, the analyses reported on in this dissertation use digital methods, as will be elaborated on below.

This dissertation also borrows from previous research with regard to the type of corpus that is used to study similarities and differences between fiction intended for different ages. Comparing

texts with a different intended readership but written by the same author limits the observation of discrepancies based on the individual author (Anderson, Shavit *Poetics*, and Van Lierop-Debrauwer and Bastiaansen-Harks). Although the term 'crosswriters' in a broader sense also defines authors who address readers of different ages in a single work of literature, I limit its use throughout this dissertation to authors who address readers of different ages in separate works. Momentarily leaving aside the classification of narrower age ranges within children's literature, there are three subcategories for this type of authors. Initially, discussions about crosswriting focused on authors who start out writing adult novels, but at some point in their careers switch to producing children's literature, either exclusively (e.g., Jef Aerts) or interspersed with their initial readership (e.g., Guus Kuijer and Philip Pullman). This practice had existed long before it was designated its own term, with Lewis Carroll's *Alice* books as a canonical example, as Marilyn Apseloff points out in *They Wrote for Children Too* (1989). Her annotated bibliography presents an overview of over a hundred famous English-speaking authors who mainly published works for an adult audience but also wrote for children. Up until the 1990s, authors who moved from writing adult to children's literature tended to keep this quiet, largely due to the relatively low status of authors of children's books. Some claimed that they saw no distinction in literature based on age (Beckett *Crossover Fiction*, 2–3) while others simply denied their function as children's authors (Shavit *Poetics*, 39). Later, the term 'crosswriter' evolved to also apply to writers who first published for children and later predominantly for adults (e.g., Hilde Vandermeer, J.K. Rowling, and Rita Verschuur), as well as writers who address both target audiences alternatingly throughout their careers. David Galef uses the term "polygraphic authors" for writers who fit into this last subcategory (29). In practice, the authors under investigation in this dissertation mainly belong to the latter category as well as to a fourth category: that of authors who write for young readers of different ages without publishing any adult literature (e.g., Anna Woltz).

1.1.2 Construction of age

In this dissertation, I study the influence the age of the intended reader, a concept which was set up in the previous section, has on the construction of age and this at two different levels, which are introduced below. To meet the first research objective (to gain deeper understanding of the construction of age in literature for young readers), the individual chapters of this dissertation address two main research questions, each operating at a distinct textual level. Chapters two through six focus on formal features to answer the question:

RQ 1.1 How does the age of the intended reader affect the construction of age *by* a work of fiction?

This question is concerned exclusively with the age of the intended reader and its interaction with the characteristics of texts, primarily in terms of form, or textual complexity, and content, or trending topics. What does a text reveal about the expectations adult producers of a text have of readers of specific ages? As indicated in the previous section, these decisions do not always lie exclusively with the author of a text, as there is possible interference from the publisher or other agent in the process between writing and reading. For example, Jacqueline Wilson was once asked by her editor to change a passage that featured one of her characters sniffing glue to prevent younger people who were unaware of this alarming trend from participating in it themselves

(*Interview with Jacqueline Wilson*), while a toilet scene and sex scene were adapted when Bart Moeyaert published his debut (Joosen, 'Writing when Young').

Although studies of the formal aspects of adult literature, especially those conducted within the field of stylometry, the quantitative study of writing style, start from the premise that an author's writing style is unique and manifests itself subconsciously (see Holmes 'Evolution', and Jockers), some authors of children's literature do consciously adapt the way they write for younger readers. From early on in children's literature research, it was established that authors use "a very informal and intimate approach" when addressing their child audience by putting themselves "on the level of the child", and writing in a conversational, personal and friendly manner (Guttery, 208). However, as established in the previous section, some authors deny that they adapt the content or writing style of their works according to their readership. The analyses that answer the first research question do not attempt to make statements about how consciously or unconsciously certain choices are made, but rather try to identify the differences in style and content and relate them with the age of the intended reader.

To answer this research question, the unannotated texts of the corpus are studied in their entirety. Because pre-processing the texts involves relatively little time and human interaction is limited (see section 1.2.1), it is feasible to include the full corpus in the analyses (see chapter 4). Furthermore, the texts are reduced to sets of numbers, including, for example, word frequencies, average sentence length, or other easily gauged surface features. Outsourcing the reading process itself to the computer and basing human interpretation on the data it generates rather than on the original text, is a common method of distant reading (Kestemont and Herman, 8). Unlike the analyses covered by research question 1.2, which use digital tools to try to look for implicit or explicit age ideologies in the text, the focus of research question 1.1 is descriptive in nature and sticks to the text's surface, also known as 'surface reading' (Kestemont and Herman). In this dissertation, I adhere to the understanding of surface as "the location of patterns that exists within and across texts" (Best and Marcus, 11). The discussions in chapters two through six rely on the comparison of sets of numerical data to reveal these patterns and in doing so contribute to the practice of readership attribution, a term coined in chapter four as a parallel to authorship attribution, an area of research within the field of stylometry. While authorship attribution aims to attribute texts of unknown or disputed authorship to their actual author, based on quantitatively measured linguistic evidence (see Holmes 'Authorship', Hoover 'Statistical', and Koppel et al. 'Author Gender'), readership attribution examines the correlation of the age of the intended reader with style, following the definition of Herrmann et al. as "a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" ('Revisiting', 44), to gain a deeper understanding of differences in textual features between works of fiction based on the age of the intended reader.

This dissertation considers not only the construction of age on the level of the intended reader, but also on the level of fictional characters. Chapters five through seven delve deeper into the corpus to answers the question:

RQ 1.2 How does the age of the intended reader affect the construction of age *in* a work of fiction?

The analyses in these chapters study meaning in the text rather than its formal features. The texts are thus not converted to numbers but to semantic units, i.e. fragments of speech, word classes,

and passages thematising age. Instead of analysing the texts in their entirety, this approach focusses on the representation of characters of different ages. Character-building is generally considered to be a process that authors navigate in a conscious manner, whether it is by building them trait by trait or discovering them while writing, as David Almond experienced while writing the title character of *My Name is Mina* (Almond, *Interview with David Almond*). Furthermore, as is especially the case in children's literature, fictional characters "serve as ideological […] vehicles" (Nikolajeva *Rhetoric*, x) that reflect the norms which function as expectations about people's appearance and behaviour within a particular socio-cultural context (Radl, 758). When these expectations are based on the age of an individual, they are referred to as 'age norms'. The dominant understanding of age in Western societies as numerical leads to normative ideas about different ages tied, for example for children, to levels of knowledge and proficiency (Morrow). Children's literature frequently portrays age-related expectations, roles, and behaviours that align with these prevailing societal attitudes and values.

Fictional characters in books for young readers are often constructed to embody and reinforce these same age norms, presenting idealised or normalised representations of specific ages. This is especially true for child or young adult characters, as even a small age difference can result in significant variations in the expected skills and abilities assigned to them, unlike adult characters (Nikolajeva *Rhetoric*, 271). The process of assigning age-related attributes to fictional characters mirrors societal expectations, thus contributing to the establishment of age norms in fiction. Part of the way in which implicit age norms are manifested in children's literature is through the semantic prosodies of words, their "tendencies to call certain associations to mind" (Knowles and Malmkjaer, 70). Knowles and Malmkjaer argue that by frequently putting two words together in a text, also called 'collocation', writers can implicitly influence their readers to draw associations between, for example, "characters and phenomena" in the narrative (71). A large-scale study of the traits assigned to characters of specific ages helps to gain insights into the broader cultural and social beliefs surrounding age, as well as the implicit or explicit messages conveyed to readers of different ages about their roles and identities within society. Chapters five to seven of this dissertation adopt some of the strategies for characterisation Maria Nikolajeva discusses in *The Rhetoric of Character in Children's Literature* (2002), ranging from implicit to explicit. To supplement the latter, chapter five discusses metareflections, which are statements made about age in general, without referring to a specific character or group of characters.

In the 1970s, a shift emerged in Western children's literature studies from research focusing on plot to an emphasis on the importance of characterisation. Following this shift, much attention has been paid to the study of fictional characters (Nikolajeva 'Narrative Theory', 168). Several studies point to differences in characterisation between books for children and for adults, which, according to John Stephens, are mainly to be found in "the way in which notions of childhood are constructed through characterisation" (Stephens 'Maintaining', 184). However, hardly any comparative research has been conducted on works intended for different ages within literature for younger people. What are the similarities and differences in the age ideologies that are embedded in characterisation in books for young children versus adolescents? While I agree that the process of characterisation is located in the space between the author's brain and the reader's brain and relies on a wide variety of strategies (e.g., Ryan's principle of minimal departure), the analyses in this thesis follow a bottom-up approach and focus only on the information contained in the text. The sources of this information are the "textually explicit ascription of properties to a character" as well as "inferences that can be drawn from textual cues", two of

the three main sources that Fotis Jannidis identifies in his discussion of fictional characters ('Character', n.p.). The third source of characterisation, which relies on references to the real world outside of the narrative, is not considered in this dissertation because it departs from the actual text. This dissertation thus adheres to a structuralist and semiotic paradigm of studying characterisation, as it acknowledges that characters are not a direct representation of human beings and is interested mainly in how the element of age is constructed in the text (Eder et al., 5). To collect the relevant data from a large set of texts, annotated material with a focus on age is used to enable this part of the research. As the annotation is a time-consuming process (see section 1.2.2), chapters five through seven report on only a part of the corpus; the oeuvres of Joke van Leeuwen, of David Almond, and the life stage of older adulthood, respectively.

Previous studies on age in children's literature are mainly limited to discussing a specific life stage, typically because traditional close reading is a time-consuming method. Furthermore, for a long time, the study of children's literature focused mainly on the construction of characters close to the intended reader in terms of age, namely children and adolescents. Research on the construction of the child in children's literature has produced a large variety of studies, ranging from historical overviews of child images (see Gavin, Sánchez-Eppler, and Pifer), to the (im)possibility of adult writers constructing child characters (see Beauvais, Hollindale *Childness*, Nikolajeva 'Neuroscience', and Rose), and how constructions of childhood influence the form and discussion of children's books (see Immel). Adulthood, however, was brought into consideration nearly exclusively in the context of the production and reception processes of children's literature (Joosen *Adulthood*, 6). Other studies that discuss the construction of adulthood do so to study the dynamics between characters in this stage of life and child characters, primarily to gain a better understanding of the latter. More recently, adult characters as well as the construction of intergenerational relationships, which studies the interaction between characters of different ages rather than regarding that relationship as being an element in the characterisation of either side, have received more attention. It is still a trend in children's literature studies, however, to consider adult and older adult characters in terms of the role they fulfil as parents or grandparents (see Apseloff 'Grandparents', Beland and Mills, Crew, Henneberg, Janelli, and Joosen 'Grandparents').

The study of age, life stages, and intergenerational relationships in a broader sense falls within the research area of age studies, an interdisciplinary field in which researchers of children's literature, among others, regularly set foot. Branching off from gerontology in the 1970s, it was initially focused on older adulthood. The long-standing unfavourable view on old age, which Simone de Beauvoir discusses in her radical book *La Vieillesse* (1970, literal translation: *The Old Age*), triggered several scholars to create a field of research to challenge negative stereotypes and affirm the individuality and agency of people in this stage of life (Segal). It is thus no surprise that 'ageism', one of the key terms in age studies, for a long time was used to describe "discrimination against older people on grounds of age" and remains to be used regularly in this narrow sense (Bytheway, 338). The rise in awareness of ageism resulted in several scholars addressing various ageist tropes, both in children's literature and beyond, including the decline narrative (Gullette, *Agewise*), the image of the infantilised senior (Hockey and James), and the stereotype of the wise old mentor (Henneberg, and Woodward 'Wisdom'). The common issue raised by these researchers is that ageist stereotypes perpetuate a social imagery which ignores the diversity of individual experiences of older age specifically, and each phase of the life course in general. While certain scholars have introduced new terminology to refer to discrimination

towards other ages, such as Sylvia Henneberg's 'reverse ageism' ('Crones', 121) and Jack Westman's 'juvenile ageism' to refer to discrimination of youth, this dissertation follows a broader definition of ageism that does not target a specific life stage as being the sole victim of ageist ideologies. This shift in perspective on the basis of age is influenced by age norms that are naturalised in society. As established in the general introduction of this dissertation, literature, and specifically fiction intended for younger readers, provides excellent material to study these kinds of implicit ideologies. Large-scale analyses of the construction of age *in* fiction, by looking both at characterisation and metareflections, will contribute to a better understanding of age norms directed towards all stages of life as well as uncover possible differences according to the age of the intended reader.

1.1.3 Children's literature and digital humanities

Traditionally, children's literature research, and indeed literature in general, is mainly conducted on a small set of texts. Due to the limitations of the human researcher, it is not feasible to process more than a handful of works at the same time, especially taking into account the time constraints imposed by research projects. In recent years, however, advances in computational technologies have enabled a shift from "close" to "distant reading" (Moretti). Instead of researchers reading a single text themselves very attentively, what Matthew Jockers calls "microanalysis", they employ computer programmes to zoom out or perform "macroanalyses", such as recognising patterns and detecting outliers in a large number of texts. The wide variety of publications, educational programmes (see Fitzsimmons and Alteri), and research projects that have appeared in the past decade prove that the field of digital humanities heralds an important next step within literary research. Whereas in 1990 it seemed "impossible to compare the syntax of all fictional child speech with all fictional adult speech" (Hurst, 17), in the 2020s this is a much more plausible avenue of research. Obviously not all fiction ever written is openly available for digital analysis today (because of intellectual rights restrictions), but almost all books that are recently published are available in a digital format and huge efforts have been made to digitise historical fiction. The number of texts that are available in online collections to be used for digital analyses is unprecedented.

The field of digital humanities does not only provide tools for quantitative research, or studying many texts simultaneously, but also to uncover linguistic aspects that are not usually the point of interest to more traditional literary research. A popular method to conduct stylometric analyses, for example, relies on the distribution of function words. Research into authorship attribution often compares a limited set of texts with established authorship to one anonymous text. Patrick Juola's famous study that uncovered J.K. Rowling as the author behind the detective novel published under the pseudonym Robert Galbraith included only five texts: the disputed detective novel, one adult novel written under Rowling's own name, and three books written by other British female crime novelists. The need for computational tools in Juola's study arises not from the size of the corpus, but from the nature and variety of features that are examined to determine similarity between the five texts, as there were "far too many for any human analyst to keep track of" (Juola, n.p.). Another example of the use of computational tools on a small corpus is Paul Martin Eve's study of just one text – David Mitchell's novel *Cloud Atlas* – including a stylometric analysis based on relative word and part-of-speech frequencies of the novel's individual chapters. In addition to the benefit of a computer 'seeing' patterns that could not be uncovered through a close reading of the material, results of a digital analysis might also be less

influenced by the subjectivity of a human researcher. However, human intervention is not completely excluded in a distant reading approach; in addition to developing software, for certain analyses the digital researcher remains tasked with setting parameters or interpreting output.

The growth of the field has introduced databases and digital tools for research, but distant reading and quantitative methods have faced criticism for their potential disregard for qualitative interpretation, as well as for their sometimes poor application of quantitative approaches. Notwithstanding the potential for digital analyses to reduce human intervention, issues of reliability and reproducibility have been central to the critique of computational text analysis in the digital humanities (see Schöch et al.). One influential critique that addresses this issue came from literary scholar Nan Z. Da, who specifically targeted the use of machine learning in literary studies. Da's publication sparked significant attention as she highlighted the "fundamental mismatch between the statistical tools that are used and the objects to which they are applied" (601). She also argued that quantifying analyses often reduce literature to numerical relations, limiting the nuanced interpretative work inherent to humanities methodologies to quantitative comparisons (see also Drucker). Furthermore, Da's critique raised concerns about the perceived homogeneity of methods in quantitative literary studies and their limited capacity to contribute to literary-critical or historical insights (Bode). Her paper elicited strong opposition from several scholars, including those whose research was the target of her criticism. Many of them countered the misconstruction in Da's critique of the role of digital methods, which she said seek to replace interpretative work in the humanities, whereas in reality, digital methods are intended to supplement and expand the methodological repertoire of the humanities (see Jannidis 'Perceived Complexity', and Herrmann et al. 'Response').

While it is no longer the case today, digital methods to study literature, when they emerged, were mainly defined in contrast to traditional literary research techniques. Digital text analyses would do everything that close reading could not but would also not be concerned with generating results that could already be obtained via different means. Despite warnings that "the 'literariness' of texts is easily destroyed" when "translating words into numbers and interpreting the numbers" (Potter, 415), this dichotomy between both approaches, close reading or microanalysis on the one hand and distant reading or macroanalysis on the other, has long been perpetuated. Putting these two types of methodology in direct opposition provoked criticism of digital humanities from many literary critics long before Da's controversial study (see Smeets), although neither Moretti or Jockers intended this new method to replace more traditional ones. Several other researchers agree with them and argue in favour of complementing one with the other (see Jänicke et al., Eve, and Underwood). The analyses included in this dissertation also try to close the gap between close and distant reading. For example, in chapter five, the final part of the analysis involves a return to the text. Here, metareflections about age are digitally extracted from the text but are interpreted using close reading strategies. By combining different types of analyses, from both close and distant reading perspectives, this chapter paints a comprehensive picture of the construction of age in the oeuvre of Dutch author Joke van Leeuwen. Chapter six is part of a multi-method approach to studying age in the oeuvre of British author David Almond. The trends identified through digital text analyses are linked to, and informed by, insights obtained via various other methods, including life writing studies, cognitive narratology, reader-response research, and media studies. The interaction between 'new' computational analyses and 'traditional' methods can take different forms. In addition to juxtaposing both techniques, researchers can search a large corpus for striking observations resulting from the study of a single

text. Or, vice versa, in a single text they can examine outliers that a distant reading of a set of texts has revealed (see Thomas 'Long Novels').

Despite the promising new opportunities for literary research enabled by digital humanities, children's literature research seemed to be hesitant to adopt a methodology informed by digital text analysis. In 2017, using computational tools was not yet seen as a common way of studying children's literature, as is evident from the structure of the *Edinburgh Companion to Children's Literature* (Beauvais and Nikolajeva), an edited volume which, in its first part, includes chapters on contemporary directions in children's literature scholarship. Eugene Giddens' chapter on 'Distant Reading and Children's Literature', however, appears in a section titled 'Unmapped Territories'. Giddens cites that a recurring point of criticism on children's literature researchers is that they are highly selective in their choice of primary material (305). While at first glance this dissertation may contribute to this self-perpetuating canon, given that the selection of the corpus departed from popular or award-winning authors, it tries to break this critique by also considering their work that has received very little attention. While digital text analyses aid in providing an answer to the critique of a small set of works that are considered suitable for research, copyright laws have long stood in the way of digitising contemporary material to be used in literary research. In other countries, where there are exceptions to copyright for scientific research purposes, digitisation backlog formed a larger issue. Giddens mentions some studies that were conducted in the 2010s and that overcame this hurdle, either by using corpora of historical material, analysing a different medium, or focussing on a smaller corpus. However, this research remains limited to word frequencies, word collocations, and speech distribution.

In recent years, however, several researchers in the field of children's literature have begun to recognise the affordances of the computational medium and methods to answer new research questions, which were not feasible with traditional literary analyses. Additionally, obtaining digital copies of children's books is no longer an obstacle due to growing online databases, such as the Auslit digital collection of storytelling (https://www.austlit.edu.au), the Baldwin Library of Historical Children's Literature (https://baldwin.uflib.ufl.edu), and the Digital Library for Dutch Literature (https://www.dbnl.org), as well as the wide offer of the eBook market. This has led to new directions in children's literature research, including digital archives (see Blake), digital genetic criticism, the study of fan-created content (see Rowe et al.), the study of images, and text analysis (see Thomas, 'Word Frequency'). Given that each chapter in this book discusses previous studies relevant to the analyses being conducted, I prefer to sketch here a picture, which in no way should be understood as an exhaustive list, of the wide variety with which digital tools have recently been employed in the study of children's literature. *The Lion and the Unicorn*, an academic journal featuring a wide array of subjects concerning children's literature, and its upcoming special issue dedicated to the collaboration between children's literature research and digital humanities, attests to the evolution of the intersection between these fields that was considered an underexplored domain just six years ago.

In the past decade, both the use of highly accessible digital tools and more complex, self-written programmes has led to innovative research. Dutch children's literature researcher Nicole Hurkmans used the software package *Wordsmith* to compare a small corpus of texts based on basic textual features. Situated on the more technically advanced side of digital text analysis is the research of Folgert Karsdorp and Antal van den Bosch, who use topic modelling to identify motifs in a large database of Dutch folktales. These two studies demonstrate that there is also

considerable diversity in the size of the corpus of primary material studied by different researchers. As discussed above, the use of digital analysis does not always dictate that it be applied to a large set of texts. For example, Wouter Haverals and Vanessa Joosen examined the age ideologies contained in Guus Kuijer's oeuvre of 32 books. Although they use a methodology that combines close and distant reading, a stylometric analysis as well as a computationally aided examination of characters' speech in a relatively small set of books enrich their findings. A leading example of quantitative analysis can be found in the study of Anna Čermáková and Michaela Mahlberg. By applying corpus linguistic approaches to historical and contemporary children's literature, they look for an evolution in the presentation of gendered body language in a corpus totalling over 17 million words.

Computational tools offer diverse research opportunities not only concerning corpus size but also the nature of the textual material covered, encompassing studies on meta-information, textual data, and illustrations. For instance, Melanie Griffin analysed paratextual data of Newbery Medal winners, confirming assumptions regarding the award's preferences for literature intended for older children, particularly historical fiction, while also exposing limited diversity in terms of voices and stories, the predominance of women as recipients, and the conservative nature of the Medal's content with gradual and predictable changes over time. Ed Thomas focusses on textual material, deriving a list of most frequent words and collocations from a large corpus of English and translated children's literature extracted from the online *Children's Bookshelf* of Project Gutenberg and consisting of stories published between 1863 and 1913. He posits that these findings are useful not only for children's literature researchers but also for the development of English language syllabi tailored to adult learners. Moving beyond research that predominantly relies on digital tools for textual data analysis, there is a growing scholarly interest in exploring the multimodal combination of text and illustrations as a means to gain a comprehensive understanding of children's literature. In a study by John Walsh, the focus lies on the attention comic books receive within the humanities, emphasising their complexity as documents with intricate relationships between textual and visual elements. His study introduces Comic Book Markup Language (CBML), an XML vocabulary based on TEI, designed to encode and analyse comic books, providing an overview of its goals, content types, key features, encoding challenges, and future plans. Paavo van der Eecken's ongoing research at the University of Antwerp represents another promising investigation that centres on the examination of both textual and visual components within the domain of children's literature. Using digital tools for visual analysis, he studies the intersectional nature of biases in illustrated children's books with the overarching research objective of identifying biases related to age, gender, ethnicity, and social class depicted in illustrations found in Dutch children's literature spanning the period from 1800 to 1940.

Tools from the field of digital humanities are not only being used to study historical material but also to examine digitally enhanced and born-digital fiction, where the convergence of technology and narrative offers novel opportunities for exploration and analysis. Claudio Vanhees et al., for example, employ statistical software to conduct a comprehensive investigation into the frequency and desirability of multimedia hyperlinks in fiction as a means to engage young readers in a pedagogical context, shedding light on the efficacy of such digital enhancements for language learning purposes. The increased engagement of young individuals with digital modes of reading, or experiencing, narratives not only entails the utilisation of digitally enhanced texts as language learning tools, but also encompasses the development of children's book applications that amalgamate written text, animation, video, audio narration, music, and sound effects. These

innovative applications leverage technological advancements to foster immersive reading experiences and facilitate multimodal comprehension among young readers. Betty Sargeant examines one such application, delving into the transformative impact of digitisation on traditional print picture books. By analysing the educational and social consequences associated with the process of remediating the picture book, Sargeant provides valuable insights into the evolving landscape of children's literature in the digital age. Her research delves into the implications of integrating interactive multimedia elements into picture books, exploring the changing dynamics of storytelling and the potential benefits and challenges brought about by digital adaptations. Through her investigation, Sargeant seeks to contribute to a deeper understanding of the complex relationship between technology, narrative, and children's literacy development, paving the way for further exploration and advancements in the field.

This dissertation contributes to the expanding body of research situated at the intersection of children's literature studies and digital humanities. While there has been a notable increase of such investigations in recent years, the current project stands out for its originality in three distinct areas. Firstly, it incorporates age studies as a research domain to explore the presence of age norms in literature intended for readers of different ages. Secondly, several of the following chapters adopt a substantially larger corpus than is customary in children's literature studies, while simultaneously integrating digital methodologies. Thirdly, this dissertation aspires to have an influence on the scholarly examination of children's literature using digital methods that can inspire further research. These innovative facets underpin the second research objective of this thesis, which is to compile a (non-exhaustive) set of methods for digital text analysis that enables the study of age-related questions, both on the level of the intended reader and of the explicit and implicit ideology expressed in fiction. The different methods employed in this dissertation are featured in the research questions that are informed by this goal:

RQ 2.1 Can stylometry lay bare formal distinctions between texts intended for readers of different ages?

RQ 2.2 Does a digital analysis of formal features (such as sentence length and lexical diversity) lay bare distinctions between texts intended for readers of different ages?

RQ 2.3 Does topic modelling lay bare thematic distinctions between texts intended for readers of different ages?

RQ 2.4 What patterns can be observed in implicit and explicit age ideologies (represented by character distributions and metareflections respectively) in the oeuvre of crosswriter Joke van Leeuwen?

RQ 2.5 Does a syntactic parser lay bare distinctions in the construction of implicit age ideology between texts for readers of different ages?

RQ 2.6 Does a digital analysis comparing works of crosswriters reflect those authors' views on writing for readers of different ages, or do they reveal other patterns?

Building upon the research questions from both research objectives, the following section outlines the methodological approach employed to investigate these inquiries, from compiling a corpus, to the annotation process and an introduction to the more complex digital analyses.

1.2 Methodology

1.2.1 Corpus building and preprocessing

The analyses in the following sections are conducted on the corpus of the overarching CAFYR research project, either on the entire set of texts or on a subset. The works in that corpus are selected based on two of the aims of this project, namely the study of the influence of the age of the intended reader, the central theme of this dissertation, and the influence of the age of the author on the construction of age in fiction for young readers. Firstly, we look for authors who write books for different ages, for children and adults, as well as for a diverse readership within children's literature. The age of the intended reader is recorded from the following sources, in hierarchical order: explicit age markers on the book itself, age markers found in other books by the same author, the author's website, the website of the publishing house, and the 'Centraal Bestand Kinderboeken', a Dutch-language catalogue of children's literature which also contains entries for a number of English titles. To fit the second research objective, we select authors who have had long careers. Because research on the influence of the intended reader uses digital tools, applied mainly to a large amount of data, to study these books, we ended up with 27 authors, whose combined oeuvres amounts to 738 prose texts. The scope of this project is both national and international as it includes books written by 11 British, 10 Dutch and 6 Flemish authors, published between 1970 and 2020. The final selection consists of 14 female and 13 male authors with a white, black or Indian background. Table 1.1 provides an overview of the number of texts, unannotated and annotated, for each author and each language included in the corpus, as well as the total number of words per author and per language.

Language	Author	Number of texts	Number of annotated texts	Average number of words per text	Total number of words
Dutch	Jef Aerts	8	0	40.937,5	327.500
	Bibi Dumon Tak	9	6	34.650,67	311.856
	Ed Franck	40	7	22.962,73	918.509
	Martha Heesen	16	0	25.043,56	400.697
	Marjolijn Hof	10	0	23.280,5	232.805
	Guus Kuijer	33	33	27.359,58	902.866
	Bart Moeyaert	40	19	10.055,63	402.225
	Anne Provoost	11	0	32.445,64	356.902
	Daan Remmerts de Vries	27	0	40.800,19	1.101.605
	Aline Sax	16	0	37.559,69	600.955
	Simon van der Geest	7	0	36.382,43	254.677
	Hilde Vandermeeren	35	5	29.836,17	1.044.266
	Edward van de Vendel	34	5	23.214,41	789.290
	Joke van Leeuwen	22	22	22.833,05	502.327
	Rita Verschuur	26	0	26.311	684.086
	Anna Woltz	25	0	29.384,72	734.618

	total Dutch	359	97	26.643,97	9.565.184
English	David Almond	21	21	33.492,27	736.830
	Malorie Blackman	58	14	28.724,21	1.666.004
	Aidan Chambers	14	3	61.704,5	863.863
	Frank Cottrell-Boyce	13	0	40.612,15	527.958
	Anne Fine	84	43	20.009,14	1.680.768
	Neil Gaiman	17	10	58.742,76	998.627
	Jamila Gavin	29	4	26.254,21	761.372
	Jill Murphy	10	10	16.925	169.250
	Philip Pullman	28	24	61.405,71	1.719.360
	J.K. Rowling	14	12	164.356,71	2.300.994
	Jacqueline Wilson	91	0	47.437,52	4.316.814
	total English	379	141	41.535,2	15.741.840

Table 1.1: Complete corpus overview

While each of these authors has an extensive oeuvre, which includes not only children's literature, but also novels published for adults, some works are not included in the corpus to ensure the most reliable analyses of the texts. First, when analysing a large amount of textual data, it is important that the data is sufficiently uniform to reduce noise. Characters who display deviant accents in their direct speech, for instance, yield outliers when studying writing style or fictional speech. In J.K. Rowling's *Harry Potter* series, the half-giant Rubeus Hagrid uses words like 'fer,' 'yer,' 'outta' and 'bin.' In a digital analysis that I carried out on the series, these highly idiosyncratic words were incorrectly flagged as highly characteristic for adult speech when compared to that of the child characters in the books—instead of being representative for adult speech as a whole, these words are characteristic of just one adult character. The analyses included in this dissertation were not affected by individual characters and thus no texts were excluded because of this. However, entire books written in idiosyncratic language do distort the results of an analysis of a larger body of texts. In practice, this only excludes one title from the corpus, namely *The True Tale of the Monster Billy Dean* (Almond, 2011), which is discussed in more detail in chapter six. Due to David Almond's authorial decisions to change the spelling of most words in this text, but to preserve the sentence structure, including this text in analyses of sentence length might not have an effect. However, as analyses of the writing style, characterisation, and the content of characters' speech in this text would not be consistent with those of the other texts, the work is excluded from the corpus of the entire dissertation. Second, picture books or graphic novels are not included in the corpus. These works, by authors including David Almond, Neil Gaiman and Marjolein Hof, are not suitable for the analyses in this thesis for two reasons: the amount of actual text is small, which puts it out of scope for stylometric analyses, for example, and a large portion of characterisation in these works happens on a visual level (Nikolajeva and Scott). Given that this thesis focuses solely on textual data, a whole layer of information would thus be lost.

Two other types of publications that are omitted from the corpus partly due to a low word count are short stories and primers. Collections of short stories have the additional issue that the age of their intended readership is often difficult to determine. Especially for British author Neil Gaiman, who has published several collections, this paratextual information is nowhere to be found. While this is also the case for some of the works that are included in the corpus, specifi-

cally for Jamilla Gavin, short story collections seem to be elusive by nature with regards to their readership. The only exception made to this method of selection are collections that feature recurring characters and settings in all short stories, for example Almond's *Counting Stars* (2000), as the stories function in a manner that resembles chapters in a novel, while they can also be read separately. Another type of publication that is not included partly due to a low word count are first reader books. In addition to the fact that the word count of these books is similarly low to that of picture books and graphic novels, these types of publications often have to adhere to a set of strict guidelines pertaining to readability and content. Part of the analyses in this thesis examines these qualities of the texts in the corpus, but the aim is rather to look for properties that are not assigned to the text beforehand. Moreover, books for first-time readers often contain plenty illustrations to help the reader anticipate what they will read about in the text. For reasons stated above, image-heavy texts are not ideal for the analyses pertaining to characterisation.

In the period between February 2019 and May 2021, 738 books were acquired through one of three methods; digital texts were donated to us by authors, publishers, and the Digitale Bibliotheek voor de Nederlandse Letteren (DBNL, Digital Library for Dutch Literature), other books were purchased in a digital format and the remaining ones were acquired in a physical format via the library of the University of Antwerp. Not all books that meet the criteria above could be added to the corpus, as some titles were so obscure that we were unable to obtain them. The books that were acquired in physical form were digitised by interns working on the project in a process of manually scanning, running the scanned documents through optical recognition software (ABBYY Finereader) and correcting the resulting texts as well as removing unnecessary paratextual information including chapter headings and page numbers. In this format, the raw digital texts were used for stylometric analyses (chapters 2 and 6), topic modelling (chapter 3), studies of readability (chapters 3 and 5), and statistical tests (chapter 4).

1.2.2 Annotation

Little over one third of the 738 digital texts acquired during the first phase of data collection went through a second round of processing to be used as input for analyses on the construction of age *in* the text through character representation (chapter 5), syntactic parsing (chapters 6 and 7) and metareflections (chapter 5). This subcorpus of 242 texts either includes all titles of an author that are selected for the main corpus (this is the case for David Almond, Guus Kuijer, Bart Moeyaert, Jill Murphy, and Joke van Leeuwen) or a sample that is representative of their work in terms of the age of the intended reader and the decade of publication, with a minimum representation of 50% for each category (this is the case for Bibi Dumon Tak, Anne Fine, Neil Gaiman, Philip Pullman, J.K. Rowling). The purpose of the annotations is to identify fragments of the text that carry an implicit or explicit ideology of age. Chapters five through seven each include a brief explanation of the parts of the annotation method relevant to the analyses conducted in those sections. Example fragments of the final method can be found there as well as links to code and data. To avoid repetition, the discussion below will mainly focus on the evolution of the method and the changes that were made between consecutive versions. For a more hands-on explanation of how the texts are annotated, appendix A includes the annotation manual distributed amongst the students and interns who helped create the annotated subcorpus. At the beginning of 2020, this manual was moved to a Github wiki-page.[i] Important to note is that both versions

of this manual still include instructions on how to fill in the header of the document, which is automated in a later version of the method.

To add information on the implicit and explicit ideology of age to the selected text, the guidelines of the Text Encoding Initiative (TEI) were used. These guidelines are used for the representation of texts in digital form, specify encoding methods for machine-readable texts (TEI Consortium) and allows the user to add information to plain texts. It has a large set of guidelines consisting of classes, elements and attributes created to mark the structure of a text as well as specific features of interest. For example, when digitising an author's handwritten notebook, tags can be inserted in the text to indicate parts of the text that are underlined, written in the margin, or crossed out. Changes between successive versions of a text can also be emphasised. Because the TEI caters to a wide range of projects, the guidelines include a large number of classes, elements and attributes that are not relevant to an annotation in order to study the construction of age. TEI allows users to make their own schema by collecting elements either from existing modules or by adding new components, as well as by putting restrictions on attribute values.

The schema customised for the CAFYR project ensures uniformity between different annotators and minimises the risk of typos which lead to syntax errors when the texts are used as input for digital analyses. First, all elements that are irrelevant to the annotation process are removed, which means that the program used for annotating, Oxygen XML Editor, raises an error when one of these elements is used. The elements that were included in the custom schema mainly pertained to the teiHeader, which records the document's meta information: title, author, project funder, name of the script developer and annotator, date and place of the annotation, the illustrator (if relevant), the place and year of publication of the book, the author's date of birth as well as their gender and nationality, the age of the intended reader and the genre of the book. Appendix A includes a full overview of the way this data is stored in the header. For the annotation of the main body of the text, elements were again excluded from the schema, but some additions to the standard set of guidelines were also made in function of the three focal points of this phase of data collection: the speech of fictional characters, the description of fictional characters and general statements about age. Non-anthropomorphic beings were not included in the annotation, because we considered a fictional character to be "a text- or media-based figure in a storyworld, usually human or human-like" (Jannidis 'Character', n.p.). While children's books regularly include characters that do not fit into this definition (e.g., talking stuffed toys), their age is often elusive and would thus not contribute to the research questions of this dissertation. Supernatural creatures were initially included in the annotations, but were eventually not included in any of the analyses for the same reason.

A distinction is made in the annotated texts between narration and direct speech by using the <said> element. Two of the standard attributes of this element were also added to the custom schema: @direct and @who. By default, the first attribute requires a truth value, either confirming or denying that the fragment enclosed in the <said> element is an instance of direct speech. In the case of a third-person narrative, narration is tagged as indirect speech (@direct="false") and dialogue is tagged as direct speech (@direct="true"). First-person narration is commonly narrated by a homodiegetic narrator who is also a character in their own story (Herman and Vervaeck *Handbook*, 85). For these types of stories, all text is tagged as character speech. The value of the second attribute identifies the character who is speaking with a unique id (e.g., @who="jane"). In the first method for the annotation of the CAFYR corpus, which was developed

and tested at the end of 2018, the <said> element has three more attributes; to record the age or life stage of this character @age and @agecat are added to the standard set of attributes and the existing attribute @cert is used to keep track of the certainty with which their age is assigned. The values of these attributes are restricted to, respectively, numbers, the names of the life stages determined in the CAFYR age model (see appendix A4) and the values 'medium' and 'high'. The latter is used in case of an explicit mention of the age of a character, whether it is their numerical age or life stage. The @cert attribute is set to 'medium' when the age of a character is assigned based on context. For a more detailed discussion on the process of assigning age from context and the questions it raises on the diversity of age as well as the risk it runs of self-perpetuating age norms, I refer the reader to chapters four, five, and six.

After evaluating this method for annotating texts, the attributes for age and certainty were removed from the <said> element. In the second method, which was used to annotate 242 texts between May 2019 and May 2022, this information is recorded as stand-off markup, which means it is stored in a separate spreadsheet. This document also includes additional information on characters: their full name, whether they are an individual or group, their species, gender, ethnicity, and relation to the protagonist of the story. More details on the different features recorded in the character lists can be found in the annotation manual included in appendix A. The character id used in the @who attribute acts as a pointer between the character list and the annotated file. This change benefited both the efficiency and accuracy of the annotations; adding less attributes to each <said> element meant it took less time to add these to the text as well as leaving less margin for error. In addition to moving information on age from the annotated document to a separate spreadsheet, the second version of the annotation method allows for the identification of reported speech or unvoiced thoughts of a focalising character. For these instances, the existing @ana attribute was added to the customised TEI schema.

The first method to annotate descriptions of fictional characters was also evaluated in early 2019 and adapted to better fit the needs of the project as well as to improve on the accuracy of the data extracted from the annotated texts. Initially, the annotator was required to identify passages related to characterisation by enclosing them in separate <seg> elements. In accordance with the custom schema, two attributes can be included in this element: @about and @cert. The first is not included in the standard set of TEI guidelines and is added to record the numerical age or life stage of the character that is described in the tagged fragment. Parallel to the information stored in the <said> element according to the first method, the @cert attribute is included to record whether the age or life stage was explicitly mentioned in the text or inferred from the context. Some of the opening sentences of *Harry Potter and the Philosopher's Stone* annotated using this method look as follows:

> <said direct="false">Mr Dursley was the <seg about="twenties" cert="medium">director of a firm called Grunnings</seg>, which made drills. He was a <seg about="twenties" cert="medium">big, beefy man with hardly any neck</seg>, although he did have <seg about="twenties" cert="medium">a very large moustache</seg>. Mrs Dursley was <seg about="twenties" cert="medium">thin and blonde and had nearly twice the usual amount of neck</seg>, which came in very useful as she <seg about="twenties" cert="medium">spent so much of her time craning over garden fences, spying on the neighbours</seg>. (Rowling 1997)

Due to the function of the fragment above, the introduction of characters at the beginning of a story, identifying passages of characterisation is quite straightforward. However, as an evaluation of this method proved, not all instances of characterisation can be objectively determined. Inter-annotator checks uncovered many discrepancies between the work of different annotators. Additionally, the large number of elements and attributes that need to be manually added to the texts makes the process time-consuming and leaves a lot of room for error. In an effort to reduce the time needed to annotate a single text, the method was adapted to a less fine-grained approach to identifying passages of characterisation and only annotating longer fragments of text related to a character. However, this method did not lead to the expected gain in time management and consistency across individual annotators. Additionally, it led to the loss of important information on the description of characters.

A new method was devised based on the principle of coreference resolution (CR), a topic that is widely studied in computational linguistics and which aims to link all references to the same entity in a text. By identifying all references to characters, i.e. pronouns, names and noun phrases referring to characters (e.g., 'mom', 'teacher'), and documenting their age, we can extract different words that have a syntactic relation to these references and, by extension, to specific ages. While the newest advances in this system show promising results for different languages and various types of media (see a.o. Chai and Strube, Xu and Choi, Park et al., De Langhe et al.), references to characters are manually annotated in the CAFYR corpus for two reasons. First, while standard models perform well for automated information extraction in texts (Flekova and Gurevych), most models are not satisfactory when it comes to differentiating between entities that have a word in common. The opposite problem also occurs often in NER systems; ambiguous references are not always classified as belonging to the same entity. Several solutions have been suggested (see Stoffel et al.), including the recent well-performing protagonistTagger, a model developed by Łajewska and Wróblewska which combines named entity recognition (NER) with named entity disambiguation (NED). My focus on age, however, complicates the matter further and forms the second reason why the annotation was done manually: in narratives that contain flashbacks or flashforwards, or that cover longer time spans, characters appear at various ages. In a sentence where a 25-year-old character expresses that they remember being gifted a bike for their fifth birthday (e.g., "I remember that time when I got a bike."), CR would correctly identify that both instances of 'I' refer to the same person but would fail to recognise that they refer to different times in their life (also see endnote iv in chapter 6). As age often remains implicit in the text, a feature that requires the reader to infer from the context, it is unlikely that it will ever be fully automatically detectable. To solve this issue, unique character ids are created in the manual annotations for every 'version' of a character. These ids are recorded as value of the @ref attribute within an <rs> element and added to the character list along with the corresponding age or life stage. The final annotation of the fragment introduced above is:

> <said direct="false"><rs ref="vernon20s">Mr Dursley</rs> was the director of a firm called Grunnings, which made drills. <rs ref="vernon20s">He</rs> was a big, beefy <rs ref="vernon20s">man</rs> with hardly any neck, although <rs ref="vernon20s">he</rs> did have a very large moustache. <rs ref="petunia20s">Mrs Dursley</rs> was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as <rs ref="petunia20s">she</rs> spent so much of her time craning over garden fences, spying on the neighbours. (Rowling 1997)

In the first version of the annotation process, the <seg> element is used to identify meta-reflections on age as well as passages of characterisation. As this part of the method remained unchanged, the <seg> element and its custom-created @about attribute were not removed from the customised TEI schema when the first method was evaluated and updated. The attribute value corresponds to the numerical age or life stage about which the generalising statement is made.

After the annotation method was fine-tuned, a Python script was written to automatically add certain elements to the text. Similar to the annotation process, this script was also developed in several stages where each version of the programme further automates the process and lightens the task of the annotator. The first version requires the user to input the narrative perspective and the quotation style (single or double) and adds <said> elements with the correct value for @direct to the text. In case of direct speech, the @who attribute is added with a default value, which is to be replaced by the annotator.[ii] Reported speech is not indicated by quotation marks and can thus not be automatically tagged in pre-annotation, making it vulnerable to annotation mistakes. In the first version of the script, the structure of the TEI header was added, but has to be manually filled out. In a later stage, the metadata of all titles in the corpus was collected in one spreadsheet, which is loaded into the pre-annotation script to automatically fill out the header as it is added to the text. The only information that needs to be added by the annotator is their name and the final date of annotation. Based on the user's input of the language of the text, the final version of the script also adds <rs> elements to pronouns, which up until then were manually added to the text. The tagging of names and nouns referring to characters is not included in any of the versions of the script, but annotators were encouraged to use a find and replace function when encountering names of frequently occurring characters (see appendix A).

Because the annotations were made by several people over the course of three years, many of which were students at the University of Antwerp who opted to annotate texts as a part of their permanent evaluation, a different Python script was written to check the annotations for accuracy and consistency. The script checks the following items: does the file include any syntax errors? Does the file validate against the custom TEI schema? Is the name of the annotator and the final date of annotation filled out? Are there any unresolved character or speaker references, or, in other words, are there any default values left in the @who and @ref attributes? Are there any imbedded <rs> elements? How many metareflections are identified in the text? Are all character ids used in the annotated text also recorded in the character list? Are there any redundant characters included in the character list? Are all required columns in the character list filled out? In addition to providing insight into the quality of the annotation, the script lists the errors so they can efficiently be corrected. The finalised annotated documents are used to analyse character distributions (chapter 5), topic modelling (chapter 6), and syntactic parsing (chapters 6 and 7). To ensure the uniformity of the data used for the narrower case studies in chapters five and seven, the oeuvres of Joke van Leeuwen and David Almond were annotated by myself.

1.2.3 Analyses

The 738 raw text files and 242 corrected annotated texts were subjected to digital analyses from various subfields: stylometry, readability measures, topic modelling, and syntactic parsing. I make use of a Python-based programme for each analysis, written either by myself or in collaboration with colleagues from the University of Antwerp, in particular Wouter Haverals,

Pieter Fivez from the text mining centre of the University of Antwerp, Wout Dillen, and Mike Kestemont. As each of these methodologies are discussed in detail in the relevant chapters, the overview below aims to be concise and provide the following information for each method: What is the method traditionally used for? To what effect is it used in this dissertation? How does the method tie in with research questions 1.1 and 1.2? What are the limitations of using this method to contribute to answering these questions? In which chapters is it employed? What type of input does the method require? What type of output does the analysis using this method generate?

1.2.3.1 Stylometry

Authors make countless choices while writing a work of fiction; some large enough to be made consciously and be time-consuming in that they might require training and revisions; others so tiny that they happen automatically, unconsciously as part of the author's writing habits. David Holmes describes style as "a preference for one or another mode of expression" ('Literary Style', 328), suggesting that the author, with possible addition of an editorial intervention, is the only factor controlling the style of a text. However, no style is solely the product of a given author, but a composition including genre, the period in which the text was written and the individual text itself (Stephens 'Linguistics', 100). Herrmann et al. present a study focusing on stylistic analysis within the context of literary stylistics, in which they introduce an operational definition of style to support interdisciplinary empirical research. They establish correlations between style and various elements of a text, such as genre, author, theme, and topic, and assert that their definition "allows to determine the style of texts, authors, genres, periods etc. in terms of a quantitative profiling of formal features" (46). The field of computational stylometry focuses mainly on the relation between style and author and is built on the premise that the author's subconscious choices define their writing style. A quantitative study of stylistic features can thus act as a fingerprint scanner to determine the authorship of a text. Stylometric analyses have primarily been used to this effect, identifying the author of anonymous or disputed texts, better known as authorship attribution, as well as in cases of the chronology of works in the oeuvre of a single author (see Temple, and Van Hulle and Kestemont). In this dissertation, stylometric analyses are used to uncover stylistic similarities and differences between texts based on the age of the intended reader rather than the identity of the author, the time of publication, or other features as proposed by Herrmann et al. As such, chapters two and six contribute to the understanding of the construction of age by a work of fiction (RQ 1.1). While chapter two also takes into account the decade in which the texts in the corpus were published (or written, in the case study on J.K. Rowling's oeuvre), the analyses reported in both chapters do not consider the possible influence of other paratextual information which has been studied as affecting an author's writing style, including genre (Koppel et al., 'Authorship Attribution'), psychologic profile (Noecker et al.), and gender (Pennebaker).

The stylometric analyses are conducted on the raw text versions of the corpus, from which a list of most frequent words can be automatically extracted. The texts are then grouped according to similar frequency distributions of the words in these lists, from which personal pronouns are first removed to limit the influence of narrative perspective. Patterns are subsequently visualised in two types of graphs from explorative data science: hierarchical cluster analysis and principal component analysis. The principles behind these (unsupervised) visualisations are discussed further in the relevant chapters.

1.2.3.2 Readability

Already in the nineteenth century, advice on the difficulty of a text was sometimes included in the frontmatter of books for children. Formulas to objectively determine the complexity on which these reading instructions, which not yet included the grade-level classification common to contemporary children's literature, were based, only started being developed in the early 1920s (Fry). To date, calculations based on various textual features, including both measures of syntactic and semantic difficulty such as sentence length, vocabulary size, total text length, and number of difficult words, are mainly used in a pedagogical or commercial context by teachers, librarians, and booksellers. The lack of complexity of texts intended for younger readers is often cited as the main difference compared to literature for adults. One of the first studies to conduct a large-scale investigation into the difference in writing style based on some of these textual properties was conducted by Celia Catlett Anderson in her doctoral dissertation in 1984. The general assumption that Anderson used as a basis for her study, namely that style in children's literature is a simplified version of that used in books aimed at adults, was taken up by other children's literature scholars over the years. Many came to the same conclusion that texts written for children tended to have shorter sentences, most often with a single clause, shorter and less complex words, concrete imagery and a higher number of verbs (see van Lierop-Debrauwer 'Grote Gelijkenis', Beckett *Crossover Fiction*, Ewers, and Wanner et al.).

Reading complexity metrics as they have been developed by various institutions are dealt with in chapter three, while the analyses in chapters four and five deconstruct the formulae by examining some of their syntactic components in isolation. By connecting formal features to the age of the intended reader of a text, the analyses of readability measures contribute to answering research question 1.1. Like the stylometric analyses discussed above, raw text files are used as input to compute sentence length, the number of sentences with subordinate clauses, word length, and lexical diversity. In chapters four and five, the annotated files, if available, are used to calculate the ratio between the number of words in direct and indirect speech. The interpretation of the results of these analyses are based on numerical output.

1.2.3.3 Topic modelling

Unlike the two previous analyses, which examine the form of a text, topic modelling is used to reveal patterns in a text or set of texts on a semantic level. Generally trained on a large corpus of texts, traditional topic models group together words that frequently appear in each other's proximity and are thus assumed to be semantically similar. One of the most widely used models in the field of digital humanities, latent dirichlet allocation (LDA), assigns scores to each word in a document according to how likely that word belongs to a particular topic (Blei et al.). When using LDA, the researcher can set the number of topics that are to be identified in a text, and thus the number of scores that are computed for each word. Topic modelling outputs vectors containing these scores, which can be visualised in different ways to compare the distribution of topics over documents or to get insights into document similarity. The applications of this tool in natural language processing are very diverse; applying topic modelling to a text can aid data analysts in efficiently processing social media posts, emails, chats etc. in large quantities. More concrete examples include tracing the evolution of certain topics in historical documents (e.g., consecutive issues of newspapers over a long period) or automatically classifying customer reviews.

In this dissertation, topic modelling is used to look at text semantics in two different ways. The first method, employed in chapter three, fits into a more traditional way of topic modelling. Starting from a model trained on a large corpus of books to create semantic clusters of words, topics are identified in a small set of texts. Based on the frequency of these topics, the analysis looks at which occur increasingly more or less in texts intended for readers of different ages. Line graphs are used to visualise the results. Using topic modelling in this way focusses on the construction of age by the text (RQ 1.1). In chapter six, the concept of topic modelling is used in a more basic manner to compare the speech of fictional characters of different ages. The underlying principle is the same between both analyses, namely, to highlight topics, in this case individual words, that are distinct for a specific document. Instead of looking at the entire raw text, the input for this analysis are the extracted fragments of direct speech associated with two specific age ranges, for example that of child and adolescent characters. The way in which a character is reported to express themselves, for example, by whispering versus shouting, adds an extra layer to the content of their speech (Nikolajeva, "Aesthetics"). The digital method that I use in this article does not allow me to take the speech tags into account when analysing direct speech, but rather focusses on the comparison of the relative frequency of each word to study the construction of different ages in the text (RQ 1.2). Jason Kessler's tool *Scattertext* translates the output data to an interactive scatterplot, allowing readers to link back to the original text, where direct speech can be read in context. A downside to this tool is that because it does not consider this context during the analysis itself, it makes no distinction between homonyms, which can result in a misrepresentation of their relative frequency.

1.2.3.5 Syntactic parsing

Whereas the previous analysis looks at the meaning of individual words in a text, without considering the order in which these words appear or the function they perform within a sentence, with syntactic parsing, I delve into the dependencies between words. As a method of natural language processing, syntactic parsing is invested in uncovering the syntactic structure of a sentence. Statistical models aim to construct syntactic dependency graphs, or parse trees, with the highest possible accuracy based on probability measures (see Collins, McDonald et al., McClosky et al., Petrov et al., Nivre et al., and Socher et al.). Parse trees can be navigated to look for specific word types and syntactic links, or dependencies, between different words, for example, an adjective and the noun it describes. Previous research has utilised syntactic parsing for a number of text mining purposes, including the study of characterisation on a large scale (see Bamman et al. 'Film', and Koolen and Van Cranenburgh). Bamman et al. ('Literary') used this method to automatically infer character types while taking into account the influence of authorial diction. They found that narrative point of view significantly influences the representation of character. To solve this issue, they suggest future research to include point of view as an effect analogous to authorship. Although narrative perspective is taken into account in the stylometric analyses in chapters two and six by removing words directly influenced by it (such as personal pronouns and different conjugations of verbs), the syntactic parser developed for this dissertation does not distinguish between narratives in first person or third person.

Chapters six and seven in this dissertation draw inspiration from the studies discussed above and use syntactic parsing to look at characterisation as represented through verbs and corresponding adverbs, adjectives, and possessions in books for younger and adult readers. The analyses in these chapters are performed on the annotated texts in the corpus and uses the added

information on character disambiguation to extract words associated to characters of specific ages from the text. As an alternative to manual annotation, Underwood et al. use the BookNLP pipeline to trace the representation of fictional characters across a large corpus. This pipeline links proper names to the corresponding character and identifies words associated with this character, including actions they undertake or undergo, adjectives to describe them and nouns that they govern. While the same word types are of interest in this dissertation, I make use of a manually annotated corpus and a self-developed script rather than employ BookNLP for two reasons. First is the matter of the changing age of characters throughout a narrative discussed in section 1.2.2 on the annotation process. Secondly, BookNLP overlooks many references to characters, for example when nicknames are used or when the narrative refers to a character by their function in the story (e.g., 'neighbour') instead of their name. While being less time-efficient, this method results in a more accurate representation of the construction of age because all references to fictional characters are identified. The frequency lists output by the parser are thus longer and better reflect the implicit age ideology through characterisation, contributing to answering research question 1.2. While my analyses do not consider narrative point of view, as suggested by Bamman et al. ('Literary'), they do address the risk of interpreting words incorrectly by taking them out of their context. For this reason, adverbs and adjectives associated with the verbs, possessions and adjectives that are being studied are also extracted from the texts.

The four analyses discussed above are the main contributions to answering research questions 1.1 and 1.2. Some of the chapters contain additional analyses, such as the use of statistical models in chapter four and an examination of the literary representation of characters of different ages in chapter five. Metareflections, as explicit expressions of age ideology, are digitally annotated and extracted from the texts and discussed using insights from a more conventional close reading technique in chapter five.

1.3 Overview of dissertation

Apart from the introductory and concluding chapters, the six chapters in this dissertation are based on articles that, with only one exception, have been published in peer-reviewed academic journals between 2021 and 2023. As of July 2023, chapter six has been published with Routledge, as part of an edited volume featuring research by all team members of the CAFYR research project on the construction of age in the oeuvre of British author David Almond. Rather than presenting the chapters according to the order in which they were written and published, the structure of the main body of this dissertation is based on the first research objective, stated in section 1.1.2, of gaining deeper understanding of the construction of age in literature for young readers. Chapters two to six provide an answer to RQ 1.1:

RQ 1.1 How does the age of the intended reader affect the construction of age *by* a work of fiction?

The first three chapters focus solely on this research question by analysing parts of the un-annotated corpus on a formal and semiotic level while taking into account paratextual information on the age of the intended reader (see section 1.1.1). Chapters five and six supplement this analysis with an investigation of the representation of age in the text, both as portrayed by fictional characters and by general statements on the topic. For these analyses, data

is extracted from the annotated subcorpus. These chapters not only answer the research question above, but, together with chapter seven, also the following:

RQ 1.2 How does the age of the intended reader affect the construction of age *in* a work of fiction?

Thus, the two research questions do not address completely unrelated concepts; rather, two chapters form a gradual transition from one concept to the other. While most of the six chapters have a distinct methodology, they all interact to respond to the second research objective, to compile a (non-exhaustive) set of methods for digital texts analysis that enables the study of age-related questions, both on the level of the intended reader and of the explicit and implicit ideology expressed in fiction. The publications they are based on were written to answer the following research questions:

RQ 2.1 Can stylometry lay bare formal distinctions between texts intended for readers of different ages? (Chapters 2 and 6)

RQ 2.2 Does a digital analysis of formal features (such as sentence length and lexical diversity) lay bare distinctions between texts intended for readers of different ages? (Chapters 3, 4, and 5)

RQ 2.3 Does topic modelling lay bare thematic distinctions between texts intended for readers of different ages? (Chapter 3)

RQ 2.4 What patterns can be observed in implicit and explicit age ideologies (represented by character distributions and metareflections respectively) in the oeuvre of crosswriter Joke van Leeuwen? (Chapter 5)

RQ 2.5 Does a syntactic parser lay bare distinctions in the construction of implicit age ideology between texts for readers of different ages? (Chapters 6 and 7)

RQ 2.6 Does a digital analysis comparing works of crosswriters reflect those authors' views on writing for readers of different ages, or do they reveal other patterns? (Chapters 2, 5, and 6)

Each chapter opens with a brief discussion of the types of digital analyses that are employed as well as how the individual chapters are situated against each other.

[i] https://github.com/lgeybels/CAFYR-wiki/wiki

[ii] This part of the script was based on a Jupyter notebook written by Mike Kestemont (https://github.com/mikekestemont/potter).

Chapter 2 A Style for Every Age

A Stylometric Inquiry into Crosswriters for Children, Adolescents and Adults[1]

Abstract – In the field of children's literature studies, much attention has been devoted to investigating differences between children's and adult literature. Works of crosswriters, authors who write for both readerships in different works, are an excellent source for this research. This article applies stylometry, the computational method of analysing style, to the oeuvres of 10 Dutch and English crosswriters to trace potential differences in their individual style and similarities between the authors. The analyses also take into account the age of the intended reader (as listed in the paratext) and the publication date, to study the influence these aspects have on writing style. Four case studies zoom in on a specific author or age category of the intended readership to study general tendencies as well as outliers. The results from the stylometric analyses are complemented with peritextual information about the author's view on style and writing for readerships of different ages. The main conclusion drawn from the case studies is that the style of the texts usually correlates more strongly with the age of the intended reader than with the time period in which it was written. Young adult literature clusters more closely with adult literature. The style associated with a younger readership is distinct in the oeuvres of most authors studied in this article and even transcends the differences between authors.

Context – Chapter two studies formal features of a large portion of the CAFYR corpus in relation to paratextual information about publication date and intended readership. The research reported here contributes to answering RQ 1.1 (how does the age of the intended reader affect the construction of age *by* a work of fiction?). Using stylometric analyses on unannotated texts, this chapter also addresses RQ 2.1 (can stylometry lay bare formal distinctions between texts intended for readers of different ages?). The analyses are performed on a large portion of the CAFYR corpus; the oeuvres of ten crosswriters, accounting for little over one third of the 738 total texts, are studied this chapter. The stylometric method employed in this chapter also informs the analyses in chapter six, where writing style is studied alongside topics and characterisation to get a greater understanding of the construction of age in the oeuvre of a single author, Joke van Leeuwen. As the research project *Constructing Age For Young Readers* focuses on the role of age in the construction of children's literature on the levels of the age of the author, the age of the intended reader, and the age of the real reader, the paratextual features included in the analyses in this chapter are rather limited and other information that might have an effect on the writing style of a text, such as genre and the author's gender, are not taken into account. The discussion of the results does link distant reading to more traditional methods of close reading and takes into account paratextual information. For example, the authors' individual views on writing for different ages are considered when examining outliers in the data, which ties back to RQ 2.6 (does a digital analysis comparing works of crosswriters reflect those authors' views on writing for readers of different ages, or do they reveal other patterns?). This chapter does lack a discus-

[1] This chapter is based on: Haverals, Wouter, Lindsey Geybels and Vanessa Joosen. 'A Style for Every Age: A Stylometric Inquiry into Crosswriters for Children, Adolescents and Adults'. *Language and Literature*, vol. 31, no. 1, 2022, pp. 62–84.

sion of the research on the effect of the age of the intended reader on the entire Dutch-language part of the corpus, while such a study yielded interesting results for the English-language part. This gap is partly filled in chapter four, where formal elements are examined statistically across both subcorpora.

2.1 Introduction

In children's literature studies, the similarities and differences between children's and adult literature have received ample attention. Several studies have investigated narrative elements such as style, genre and content (e.g., Appleyard, Stephens 'Maintaining', Talley). As early as 1984, Celia Catlett Anderson studied the style of children's literature alongside books published for adults. She argues that comparing passages of so-called 'crosswriters', authors who write for readers of various ages in separate works,[i] is the best method to identify differences between children's and adult literature: it "eliminates the problem of differing idiolects" and reduces the impact of content, theme and genre (40).

Barbara Wall contends that "adults, whether or not they are speaking ironically, speak differently in fiction when they are aware that they are addressing children. Such subtleties of address define a children's book" (2). Scholars who have used narrative analysis to study the work of crosswriters confirm that the language and content differ. In this article, we supplement their insights, as well as authors' own views on writing for different audiences, with a stylometric approach. Our method relies on tools developed in the field of digital humanities to measure similarities and differences in style and get a better understanding of the way crosswriters address various audiences. To this aim, we have subjected large parts of the oeuvres of 10 British, Dutch and Flemish contemporary crosswriters to a stylometric analysis: David Almond, Anne Fine, Ed Franck, Neil Gaiman, Guus Kuijer, Bart Moeyaert, Philip Pullman, J.K. Rowling, Joke van Leeuwen and Hilde Vandermeeren. In essence, the stylometric method is very similar to that of traditional literary stylistics; the difference being that modern, computer-aided stylometry aims to steer away from the researcher's deliberate attention on certain stylistic features, shifting to more "backgrounded features of authorial style" (McIntyre and Walker, 65). Consequentially, stylometry tends to be less concerned with the analysis of meaning-carrying words or patterns, but focuses more strongly on the exposure and comparison of certain distinctive, grammatical features (Mahlberg and Wiegand, 308). Previous research in this area has already led to impressive results, for instance, in the identification of anonymous or pseudonymous authors and in periodising authors' oeuvres. In this article, we argue that stylometry can also contribute to a better understanding of the way crosswriters' style varies, depending on the intended readership of the books.

We approach the stylometric analysis with the following research questions: do crosswriters' works tend towards a stylistic similarity according to the time period in which they were written, or is the age of the intended audience a more determining factor? For determining the age of this intended readership, we rely on publishers' information and library catalogues. We realise that this rating may not solely be the result of the author's decision, and that the publisher and other mediators are also involved in setting the age of the intended readership. Moreover, while some authors display a keen awareness of their readership during the writing process, for others, considerations about readership only come in hindsight. Since authors rarely express themselves in detail about the intended audience of their books, we relied on information from publishers and libraries, since such indicators were available for all titles. Such paratextual age recommendations are usually meant as a minimal age, since children's books can have multiple addressees, including adults (Shavit 'Double Attribution').

In addition to the distinction between children's books and adult books, we also consider a segmentation within children's literature. After all, children's literature is a broad term that encompasses works from baby books to young adult literature. We examine whether a more fine-grained segmentation of crosswriters' oeuvres is also reflected in the stylistic analysis. Are titles that are classified as adolescent fiction clustered with the authors' children's books or rather with their adult works? In analysing the material in the light of these questions, we consider individual differences between authors, recurrent trends as well as interesting outliers (books that cluster in unexpected ways). Moreover, we also have carried out stylometric analyses across authors' oeuvres, to investigate which aspect of the text is more dominant: the intended readership or an author's idiosyncratic stylistic features.

2.2 Research into crosswriters

Several scholars, including Zohar Shavit ('Double Attribution'), Sandra Beckett (*Crossover Fiction*) and Helma Van Lierop-Debrauwer ('Grote Gelijkenis') have analysed selected works by crosswriters to explore whether and how they adapt their works according to the age of their audience. Adaptation theory identifies four ways in which children's literature authors adjust their work: content, style, structure and design (Van Lierop-Debrauwer 'Grote Gelijkenis', 343). Studying these aspects in the works of crosswriters can reveal differences between literature written for children and for adults. The main finding is that texts for children have a simple style that is clearer and more concrete than those published for adults (Beckett, *Crossover Fiction*). The critics' evaluation of style often relies on the complexity of texts, measured by the length of the full text, as well as its paragraphs, sentences and words. In addition, Rita Ghèsquire (*Verschijnsel*) and Nicole Hurkmans discovered that books for children often have a higher percentage of dialogue compared to narrative text.

Most research into the differences between children's and adult literature has been carried out by a detailed, close reading of a small corpus. For the purpose of exposing subtle trends in large corpora, however, close reading is not particularly suited. The mere process of reading, say, 300 novels, would already take quite some time. Any thorough analysis that ensues would also greatly challenge the limits to the storage capacity and working memory of a human researcher. Because of these limits to human performance, a quantitative approach, where literature is condensed into 'data', can help to break new ground. As a result, the vantage point from which this line of research is conducted is no longer close, but distant. First described by Franco Moretti, the phrase 'distant reading' refers to any form of reading aided by the computer. It can facilitate both the analysis of large corpora (what Matthew Jockers calls 'macroanalysis') and a quantitative reading of texts, taking into account linguistic aspects that researchers in literary studies do not usually pay attention to (such as the distribution of function words). In this respect, it is important to note that the concepts of 'close' and 'distant reading' are not mutually exclusive. Several researchers are convinced that the two reading strategies are complementary and can thus reinforce each other (see Jänicke et al., Underwood).

Distant reading methods have only rarely been applied to crosswriters, but the scarce studies that do exist demonstrate their potential. Hurkmans, for example, used the programme *Wordsmith* to investigate the relative frequencies and distribution of dialogue in the work of four Dutch-language crosswriters. Melanie Griffin conducted a computational analysis of Newbery Medal winners to identify patterns in American children's literature. These are pioneering

studies, but both contain shortcomings or disadvantages. Hurkmans' findings, for example, are based on a limited corpus of two novels per author, from which only 15 pages were selected. Griffin's observations are mainly restricted to meta-information retrieved from bibliographic records and are less informative about the actual content or stylistic features of the books.

In this article, we apply computational distant reading techniques to study the stylistic features of the works of several crossover authors. In this respect, an important element of the current research is that we do not limit ourselves to a random selection of works. Rather, for the authors under scrutiny in this article, we have collected and digitised their complete oeuvres – albeit with some minor reservations (see below) – so that a distant reading can be pursued. Additionally, we let the works speak for themselves. In our quantitative stylometric analysis, it is up to the computer to determine the distinguishing characteristics between the works; interference or manipulation on the part of the researcher in this process is limited. That being said, human interpretation is still needed to make sense of the results of the stylometric analysis, which may also spark new research questions with which to approach the books via close reading.

2.3 Stylometric method

Although stylometric methods predate the use of computers, the rise of digital humanities has stimulated an increased interest in the quantitative study of style (Holmes 'Evolution'). Particularly in the field of authorship attribution, stylometry has made its mark. Contrary to more traditional methods, computational stylometrists pay little attention to conspicuous word choices or striking syntactic structures that are assumed to be tell-tale characteristics of an author's writing style. Greater importance is attached to so-called 'function words', a limited set of highly frequent words, such as prepositions, pronouns, conjunctions, particles, and articles. As they "reflect deeply ingrained linguistic habits" (Hoover 'Statistical', 422), there is a great deal of evidence for their status as discriminators for an author's writing style (Argamon and Levitan, Kestemont). Moreover, function words are particularly useful for the study of authorial style since they hardly affect the meaning of a sentence, and their occurrence is largely independent of topic and genre (Binongo). By investigating their use and proportional distributions, several anonymously published novels have been convincingly attributed to an author. One notable study was conducted by Patrick Juola, who provided ample support to identify Robert Galbraith as being the 'nom de plume' of J.K. Rowling.

While there is considerable scholarly interest into authorship attribution, researchers have also explored the ability and ways in which a single author can adopt different writing styles. An interesting case in this respect is Plato. Various studies provide evidence for a stylistic development in the classical philosopher's prose (Lane). Some stylistic features might have been deliberately adopted, while others might have evolved rather unconsciously (Brandwood). Although scholars have made attempts at proposing a relative chronology of Plato's work based on these stylistic variations, not all periodisations are accepted. Still, with some confidence, Temple was able to identify a group of 'late' works, in which Plato also demonstrates his dexterity for stylistic variation. Another noteworthy application of stylometric methods was conducted by Dirk Van Hulle and Mike Kestemont. By studying the linguistic development in the works of Samuel Beckett, they argue for a more nuanced periodisation of Beckett's oeuvre. Interestingly, their findings are supported by statistical analyses of highly frequent function words.

In children's literature studies, stylistic analyses have mostly been conducted to study differences in translations (see Sikorska, Rudvin, Alvstad, Čermáková, Malmkjær, and Toolan). We argue that computer-aided stylometry also provides a promising strategy for an inquiry into the work of crossover writers. Before moving on to the findings, it is essential to shed some light on the process of collecting and preparing our corpus, as well as the settings and parameters that we used for our analyses. Our study focuses on the work of 10 contemporary crossover authors, of whom five write in Dutch and five in English. We selected authors with (relatively) long writing careers so that we could assure ourselves of an adequate supply of textual material and also consider periodisation. For Dutch, the choice fell on Ed Franck, Guus Kuijer, Bart Moeyaert, Hilde Vandermeeren and Joke van Leeuwen. For English, we selected David Almond, Anne Fine, Neil Gaiman, Philip Pullman and J.K. Rowling. Although we strived for exhaustiveness, some titles were omitted from the corpus. First, our study is limited to fictional works. Second, books for novice readers were excluded. Both content- and style-wise, this material differs a priori strongly from fiction for more advanced readers, let alone from adult novels. Moreover, these works do not often contain an ample number of words, which renders a stylometric analysis challenging (López- Escobedo et al.). Third, (collections of) short stories were also discarded. In part, the reason for this is similar to the one just mentioned. For example, Van Leeuwen's short story *Twee beleefde dieven* (*Two polite thieves*) consists of only 325 words. A second reason for excluding short stories has to do with the difficulty of determining their intended readership. For many of Gaiman's short stories, for example, it is unclear when they were originally written and for which audience they are intended. By excluding short stories, we aim to guarantee the corpus' uniformity. Table 2.1 provides an overview of the number of works collected – in the form of computer-readable files – per author. Additionally, the table shows how the novels are subdivided according to the age of their intended audience. This information was obtained from the paratext, primarily from the books themselves, if an age range featured in the colophon. When this was lacking, we either consulted the website of the publisher or library catalogues (e.g., Centraal Bestand Kinderboeken). Specific numerical ages were subsumed under broader life stages. To this end, we used the following age model: middle child (ages 6 to 8), late child (ages 9 to 11), young adult (ages 12 to 18), and adult (18+).

Author	Collected novels	Novels categorised by age of the intended reader			
		Middle child	Late child	Young adult	Adult
David Almond	19	4	3	11	1
Anne Fine	82	41	18	14	9
Neil Gaiman	17	4	4	1	8
Philip Pullman	25	3	6	13	3
J.K. Rowling	13	1	3	4	5
Guus Kuijer	32	8	15	4	5
Bart Moeyaert	17	2	3	10	1
Hilde Vandermeeren	34	15	10	0	9
Joke Van Leeuwen	22	5	10	2	5

Table 2.1: Corpus overview.

As highlighted above, much of today's stylometric research is based on the statistical occurrence of highly frequent words (see Burrows 'Delta', Binongo). For our analyses, we made use of the

100 most frequent words (MFW).[ii] These numbers are consistent with those used in similar stylometric research (see Stamatatos, Van Hulle and Kestemont). At the same time, the frequencies of certain personal pronouns might reveal something about the narrative point of view, rather than about an author's distinct writing style. To prevent the stylistic analyses to be affected by a text's narrative perspective, it is a common practice to disregard personal pronouns (Burrows *Computation*). This so-called practice of 'pronoun culling' is something that we carry out in our investigation as well. By way of illustration, an extraction of the 100 MFW for the English novels in our corpus yields the following list (words preceded by '#' were excluded from our analyses):

> a, about, again, all, an, and, are, as, at, away, back, be, been, before, but, by, can, could, d, did, do, don, down, even, for, from, get, go, got, had, have, #he, #her, #him, #his, how, #i, if, in, into, is, it, just, #know, like, little, ll, looked, #m, #me, more, #my, no, not, now, of, off, on, one, only, or, out, over, #re, right, #s, #said, #see, #she, so, some, t, that, the, #their, #them, then, there, #they, #think, this, #thought, through, #time, to, up, was, way, #we, well, #were, what, when, where, who, will, with, would, #you, #your

For several case studies described below, we performed a hierarchical cluster analysis (HCA). This method visually depicts the stylistic similarities and differences in a tree diagram, or a so-called dendrogram. The underlying principle is that (groups of) texts containing similar frequency distributions of the analysed MFW will be merged, resulting in a cluster. Conversely, the more branches that separate two texts from each other, the greater their stylistic difference. To expand on our investigation, we also produce scatter diagrams, which show the result of a principal components analysis (PCA). In this case, the texts under investigation are presented as dots in a two-dimensional space. This representation provides, as it were, a synthesis of the most important – hence 'principal' – variation that occurs in the analysed data (in our case MFW). The way in which a PCA should be interpreted is somewhat similar to that of the aforementioned HCA. Texts or (groups of) dots that appear in each other's proximity are stylistically more similar than (groups of) dots that find themselves at greater distance from each other (Binongo and Smith). Finally, it should be noted that the lengths of the novels in our corpus – obviously – vary. For example, Rowling's *The Order of the Phoenix* consists of roughly 250,000 words, whereas Fine's *Scaredy-Cat* only has about 4000. To prevent unequal text length from affecting the analyses, in several of the following case studies we applied a sampling technique that is very common in stylometry studies (Eder). In practice, this means that texts are chopped up into consecutive, non-overlapping samples of a fixed word length. Trailing words at the end of a text are discarded if the number is below the required threshold to form a new sample. After this procedure, the stylometric analyses are carried out for the individual samples.[iii] In the following section, we present the results of this analysis for a selection of individual authors as well as for the overall stylistic comparison for each language.

2.4 Case studies

2.4.1 Hilde Vandermeeren

Hilde Vandermeeren is a Flemish author who made her debut in 2001 with *Een vroege zomer* (*An early summer*), a book for late child readers (10+). Since then, she has published 55 titles, 25 of which fit the criteria for our stylometric analysis. In interviews, Vandermeeren often refers to her degree in psychology and her experience as a teacher and mother to describe her approach to

children's books. "I can perfectly identify with a nine- year-old child," she claimed (in Wybo, 14).[iv] She often mentions the age of the intended reader when describing her writing process, and says that she adapts the language and style accordingly, taking even 1-year differences into account (in Kortemark). Her own children have also commented on her drafts (AFT). In 2013, Vandermeeren made a switch to thrillers and soon after, she stopped writing children's books. The diversity of her oeuvre as well as her reflections on the age of the intended reader raise the hypothesis that a stylometric analysis would display a clear division between her books for young and adult readers. It also raises the question whether such an analysis also shows a more fine-grained distinction between middle and late children in the intended readership.

Given the diversity in length in Vandermeeren's work, we divided her books into samples of equal size and ran an analysis with the 100 most common words (76 after exclusion of certain content words and words related to the narrative point of view).[v] The hierarchical cluster analysis (Figure 2.1) shows an almost categorical division between works for children and works for adults. One children's book for 9-year olds, *Mijn geheime papa* (*My secret dad*) is clustered with her adult work. The PCA (Figure 2.2) shows that this book lies in the middle of a gradual, but very clear transition from children's books for middle readers (right) to late child readers (middle) to adult work (left). The only real outlier is one adult book that surfaces in the children's cluster in both the dendrogram and the PCA. As it turns out, *Moord in de wijk* (*Murder in the district*) was published in the 'Wablieft' series, a series for adults that aims to offer literature and journalism in clear, accessible language that does not exclude adults with reading difficulties. About the writing process of *Moord in de wijk*, Vandermeeren explicitly stated in an interview that "it took relatively little effort to come up with the plot and use clear language. Important for the target audience is the positive feeling that prevails afterwards. In the sense of: 'Wow, I've read a book!'" (in Sabbe, 10).[vi] Vandermeeren's alleged considerations for the intended readership are thus confirmed in the stylometric analysis, which lays bare an interesting distinction in adult readership. This evolution also supersedes periodic and generic distinctions. Vandermeeren writes mystery and detective novels for children and adults, but these are clustered according to age (with the detectives and realist fiction for children being grouped together), not genre. The stylometric analysis confirms our hypothesis that Vandermeeren adapts her writing style strongly when writing for children and adults, and that her books for children also display subtle distinctive stylistic features that can be related to age.

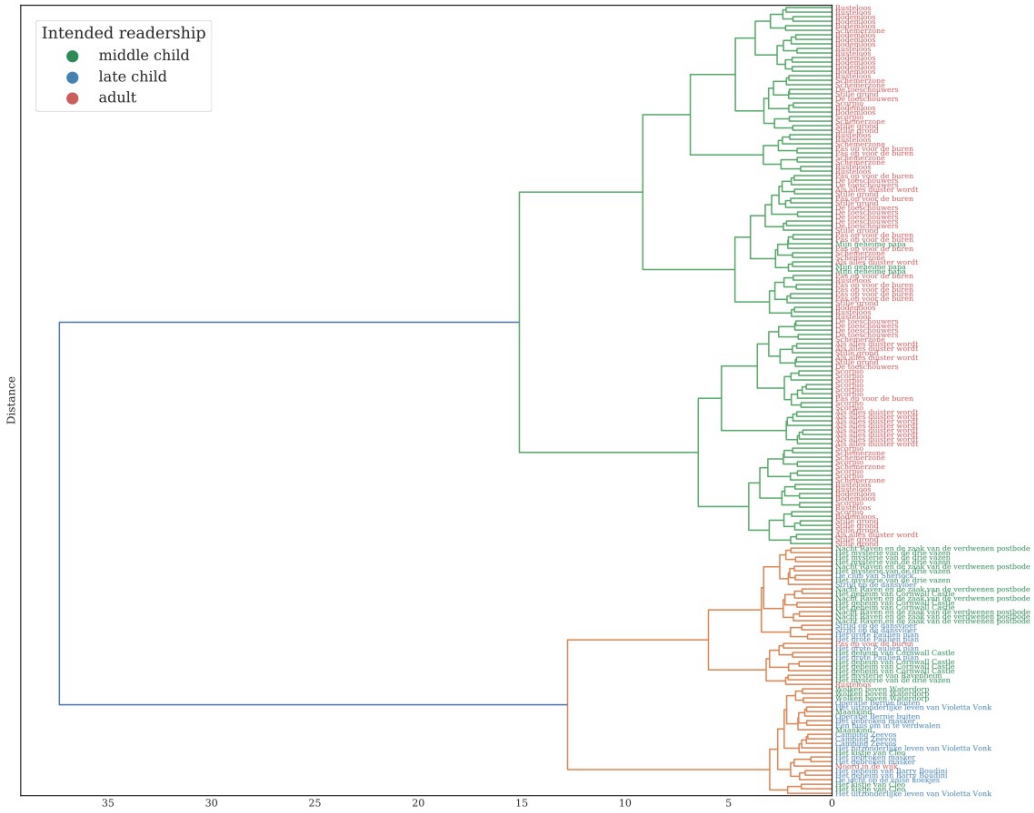


Figure 2.1: HCA for 5000-token samples of books by Vandermeeren.

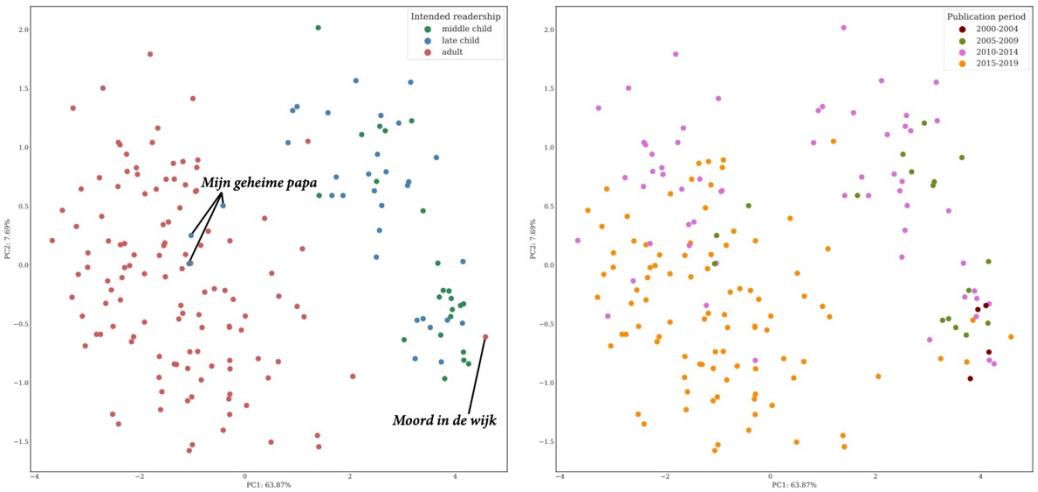


Figure 2.2: Two PCA's for 5000-token samples of books by Vandermeeren.
Left: text samples coloured according to intended readership.
Right: text samples coloured according to publication period.

2.4.2 Guus Kuijer

In the course of his prolific career, Astrid Lindgren Memorial Award winner Guus Kuijer has also dedicated himself to various audiences of intended readers. In comparison to Vandermeeren, Kuijer has made the reverse 'shift'. He made his debut in 1971 with a collection of adult short stories, *Rose, met vrome wimpers* (*Pink, with pious eyelashes*). It was not until 1975 that he

started writing children's books and earning great fame. At later stages in his writing career, he still occasionally published adult novels and also started writing for adolescents.

If we subject Kuijer's works of fiction to a stylometric investigation, we can identify similar tendencies as those observed for Vandermeeren's oeuvre, although in the case of Kuijer, they are more subtle. This might be explained by the course both authors took in their writing career; while Vandermeeren has a clear break in her works where she moved from writing for children to adult novels, after his children's literature debut, Kuijer published for both readerships alternately. Moreover, in his collection of essays *Het geminachte kind* (*The despised child*), Kuijer deplores strict boundaries between didactic children's and more literary adult literature and scolds authors who do not take the critical capacities of young readers seriously (125–36). Nevertheless, a PCA (based on the 100 MFW and for samples of 5000 words), shown in Figure 2.3, reveals that his novels do bear witness to a stylistic distribution according to intended readership. The samples of novels for adults are located mainly in the left half of the chart, while those for a younger intended audience are situated on the right. Simultaneously, the works for young adults tend more towards those for adults, while books for late children seem to be stylistically more akin to those for middle children. Regardless of this general trend, there are also some striking outliers. Kuijer's first two novels, *Het dochtertje van de wasvrouw* (*The washerwoman's daughter*, 1980) and *De man met de hamer* (*The man with the hammer*, 1975) are stylistically further apart from his later adult fiction.

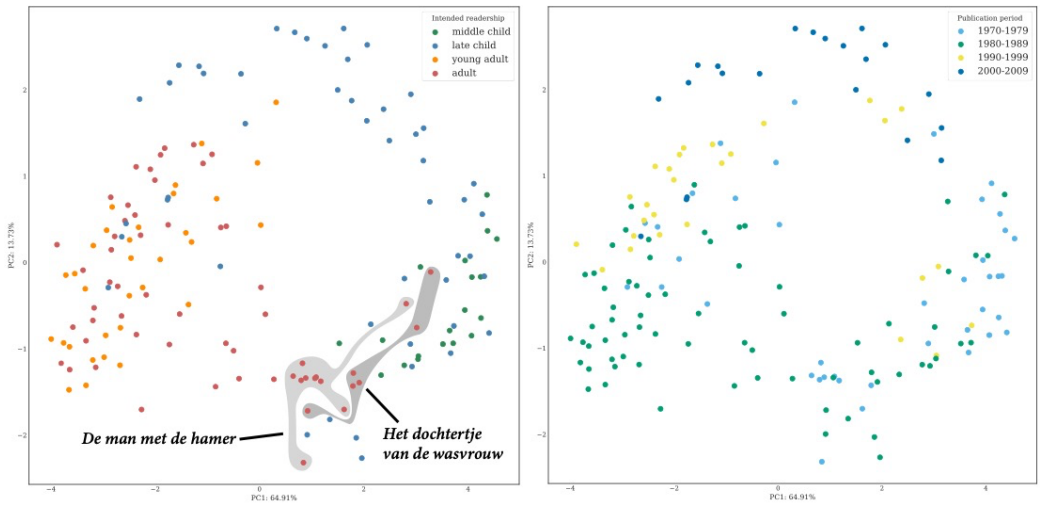


Figure 2.3: Two PCA's for 5000-token samples of books by Kuijer.
Left: text samples coloured according to intended readership.
Right: text samples coloured according to the publication period.

In the hierarchical cluster analysis (Figure 2.4), the majority of the samples of these two texts also clearly cluster together and are at a distance from the other books for adults. Interestingly, in interviews, Kuijer mentions a deliberate rejection of complicated language during the writing process of these two novels. He states that "*Het dochtertje van de wasvrouw* is an attempt to write more plainly" (in Van Den Hoven 'Kijk, Volwassenen', 48). Since he felt that his attempt had failed, he turned to children's books while completing *De man met de hamer*: "if you are going to write for children, you can't do that [using a complex style], because they see right through it" (ibid.).[vii] His frustration with literary pretentiousness in writing for adults led to a successful experience in writing for children: "Ultimately, that is where *Met de poppen gooien* (*Throwing*

dolls, 1975) came into being," (ibid.) Kuijer explains. Remarkably, both elements of this poetic statement are apparent from the PCA (Figure 2.3). First, Kuijer's apparently self-conscious attempt to draw up his first two adult novels in plain language can be observed in the PCA. The samples of both *Het dochtertje van de wasvrouw* en *De Man met de hamer* occupy a central position, in between his later adult fiction and his novels for children. This can be observed in the HCA as well (Figure 2.4), albeit that the image that this analysis presents is somewhat less delineated. Here, some samples taken from Kuijer's first adult publications are clustered with samples from his early books for children. Second, Kuijer's intent of uncomplicated writing for children can also be observed. The PCA (Figure 2.3) shows that his first books for children (e.g., all novels in the *Madelief* series) are on the far right, well removed from his adult and young adult work.

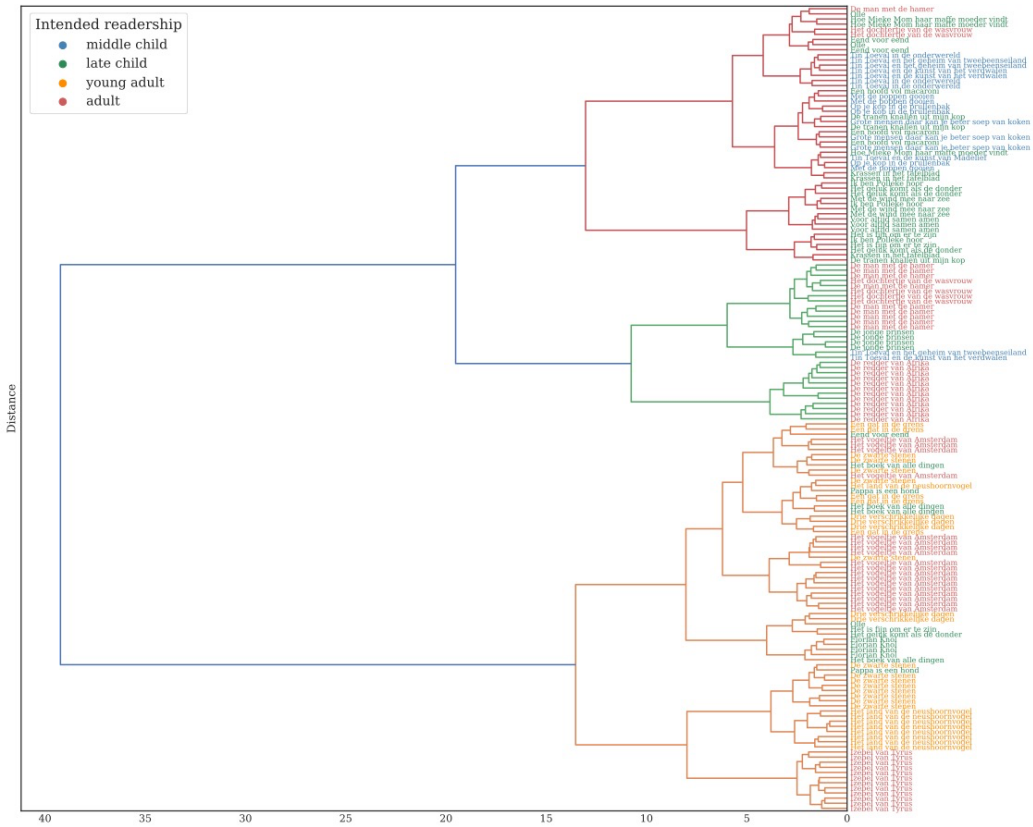


Figure 2.4: HCA for 5000-token samples of books by Kuijer.

The first principal component (PC1) in Figure 2.3 (PCA on the left) appears to confirm Kuijer's deliberate choice to adapt his style according to the age of the intended audience. However, as mentioned earlier, this analysis is based on merely 100 highly frequent function words, of which it is assumed that they are used rather unconsciously by an author (Stamatatos). In order to probe for deliberate choices, we carry out an additional Zeta test for contrastive text analysis (Burrows, 'All the Way'). This way, we are able to compare two subcorpora (selected sets of texts): Kuijer's adult and young adult fiction on the one hand, and his work for late and middle children on the other. For the calculation of Burrows' Zeta, each group is first subdivided into samples of a fixed token length (here 2500 words). Next, the appearances of all words are counted in each sample. Burrows' Zeta then returns a list of words that are statistically either preferred or avoided in each subcorpus. For Kuijer, the results of such a Zeta test are shown in Figure 2.5. Several interesting observations can be made from this chart. What certainly stands out are the verb tenses (bulleted words). In the (young) adult subcorpus, Kuijer clearly favours

the past tense, while the actions in his children's books take place in the present. Moreover, the actions in his children's books are of a more physical nature than those in his adult books. Especially distinctive for his children's books is the explicit marking of the way in which something is being said. Favoured in this respect are the verbs 'roept' ('yells'), 'fluistert' ('whispers'), 'schreeuwt' ('shouts') and 'zegt' ('says'). Finally, from Figure 2.5 we also learn that Kuijer favours the use of interjections in his children's books, such as 'hè' ('right?', 'isn't it?') and 'nou' ('well'). He rarely uses these words in his (young) adult fiction. The Zeta test thus shows that Kuijer's stylistic differentiation is not just limited to function words, but also becomes apparent in content words.

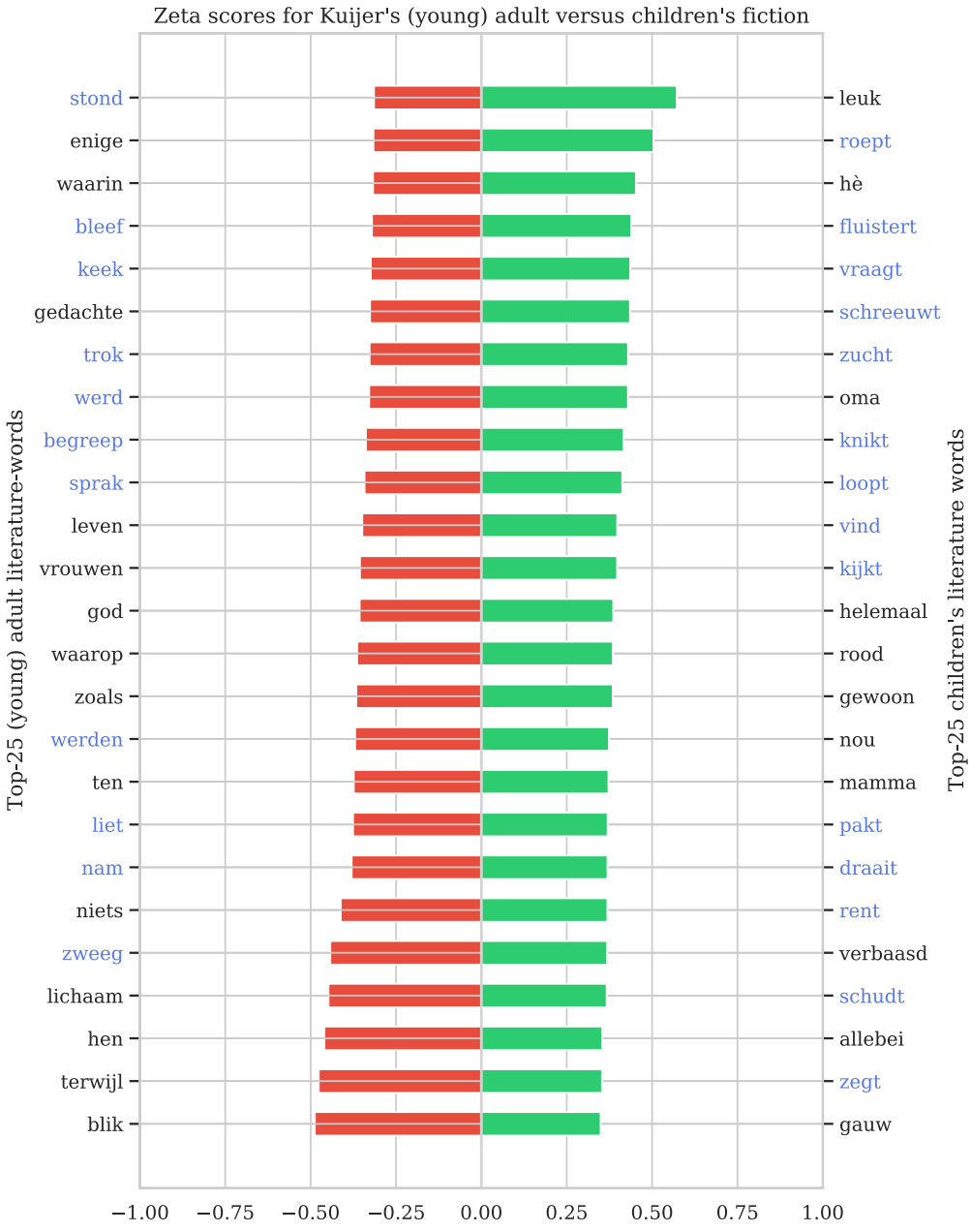


Figure 2.5: Burrow's Zeta for Kuijer's (young) adult fiction versus middle and late children's literature (on a 2500-token sample basis).

2.4.3 Corpus-wide trends for English titles

While the stylometric clustering is fairly neat for Vandermeeren and Kuijer, this is not true for all authors, as we will show. In fact, the distribution is different for each author: some oeuvres seem to distinguish themselves stylistically from each other according to the intended reader, while other analyses suggest a greater influence of the year of publication on the writing style. Despite these idiosyncrasies, two trends operate across oeuvres. First, an overlap between titles for young adults and adults can be observed. Second, the style in books published for 'middle child' readers is distinct in the oeuvres of almost all authors studied in this article. To investigate whether the stylistic features of these texts transcend the differences between authors, we have studied all authors together for each language (5 for Dutch; 5 for English), and see the same overall trends for both Dutch and English. Most analyses show a distinct cluster of titles published for 6- to 8-year olds. Since we have just covered two case studies in Dutch, we will now turn to the results for English.[viii] In total, 155 titles by Almond, Fine, Gaiman, Pullman and Rowling are included in this analysis. Because of the large number of titles, the analysis is conducted on the full, unsegmented, texts.[ix] The PCA is again generated using 100 MFW.

Figure 2.6 shows three scatter plots of English titles based on the age of the intended reader (top), the author (bottom left) and the publication period (bottom right). The last factor produces a very fuzzy image that suggests that publication periods do not play a big role in stylistic similarity. Authorship (Figure 2.6, bottom left) proves to be an impactful factor for the dispersion of titles. In particular, the works of Anne Fine and David Almond form cohesive clusters. The close proximity of works by Philip Pullman, J.K. Rowling and Neil Gaiman – all three famous for their fantasy works – suggests that genre has an impact here too, but note that Rowling's detective novels for adults and Pullman's realistic works *The White Mercedes* and *The Broken Bridge* are also part of this cluster. Within the works of Fine, the age of the intended reader is an additional influential factor, with her work for middle child readers clustering in the top and middle zone, and her work for young adults and adults clustering in the lower parts of the scatter plot. In the overall picture for intended readership (Figure 2.6, top), adult and young adult titles are mixed and located mostly at the bottom and on the right of the scatter plot, a trend that is also present in the analyses conducted on all Dutch texts in the corpus. The books categorised as 'late child' are the most dispersed, while those for the youngest age category are mostly located on the left. The vast majority of titles in this left half of the plot are written by Fine, which demonstrates that the first component (PC1) mainly picks up on the stylistic difference between Anne Fine and the other authors in the corpus. There are, however, several outliers that can be identified when looking at this cluster in more detail. There are several titles categorised as 'late child' that appear in the 'middle child' cluster. However, only one YA title ventures that far to the right of the PCA: Almond's *Klaus Vogel and the Bad Lads*. It is explicitly stated on Almond's website that this is a book for readers aged 12+ ('Klaus Vogel'). When looking at the cluster analysis (Figure 2.7), the one-dimensional representation of the PCA, Almond's *Klaus Vogel and the Bad Lads* is surrounded by books written by Fine on the bottom of the graph. An explanation for this observation is close at hand as Almond wrote this book "for struggling, reluctant and dyslexic readers" ('Klaus Vogel') and published it at Barrington Stoke. According to the motto of this publishing house, printed at the end of each title, their "books are tested for children and young people by children and young people". They focus on readable and dyslexia-friendly children's books and also published a handful of Fine's books, which might explain the stylistic proximity of *Klaus Vogel and the Bad Lads* to her work. Apart from the notable presence of a young adult title in the cluster

of ‘middle child’ books, there are two titles that are outside of this main cluster and gravitate more to the left side of the PCA: *Count Karlstein* by Pullman and *The Ickabog* by Rowling.

Figure 2.6: PCA visualisations of the distances between all books by English authors in the corpus. PCA on top shows the variation among novels according to the age of the intended reader. PCA’s at the bottom show the variation according to the author (left), and according to the publication period (right). Note: PCA: principal components analysis.

Pullman has been addressing a diverse readership for several decades. *Count Karlstein* (1982) was his first children’s book, after writing two adult novels. During his career, he published two more books in the same age category: *The Firework Maker’s Daughter* (1995) and *Clockwork* (1996). These two books are stylistically much closer to the cluster of ‘middle child’ books as visualised in Figure 2.6. This suggests that the age of the intended reader is not the only aspect influencing Pullman’s style. The author confirms this observation, as he has stated more than once that he never writes with a specific readership in mind. He believes “that children’s books

belong with the rest, in the general field" (in Lister). In the Q&As on his website, he answers the question for whom he writes, children or adults, as follows: "Myself. No-one else. If the story I write turns out to be the sort of thing that children enjoy reading, then well and good. But I don't write for children: I write books that children read. Some clever adults read them too" ('Questions'). It is unlikely that Pullman wrote his books with a specific age range in mind. There is, however, a difference between the writing process and the decisions made by the publisher, as the author emphasises himself. For example, Pullman conceived his most famous trilogy, *His Dark Materials*, as appropriate for all ages. However, it was published and marketed for young adults (Beckett *Crossover Fiction*, 118). In the stylometric analysis, we observe an evolution in Pullman's writing style, which may not be conscious or deliberate (as we saw above, authors rarely reflect on their use of function words). The style of Pullman's first children's book, *Count Karlstein*, still approaches his writing for adults. As his career developed, his works show a greater variety of stylistic features, and a larger distance appears between his texts for young adults and adults on the one hand and young children on the other. While Pullman might not actively write for a specific age, a distinct style for younger readers has crystallised in the course of his career.

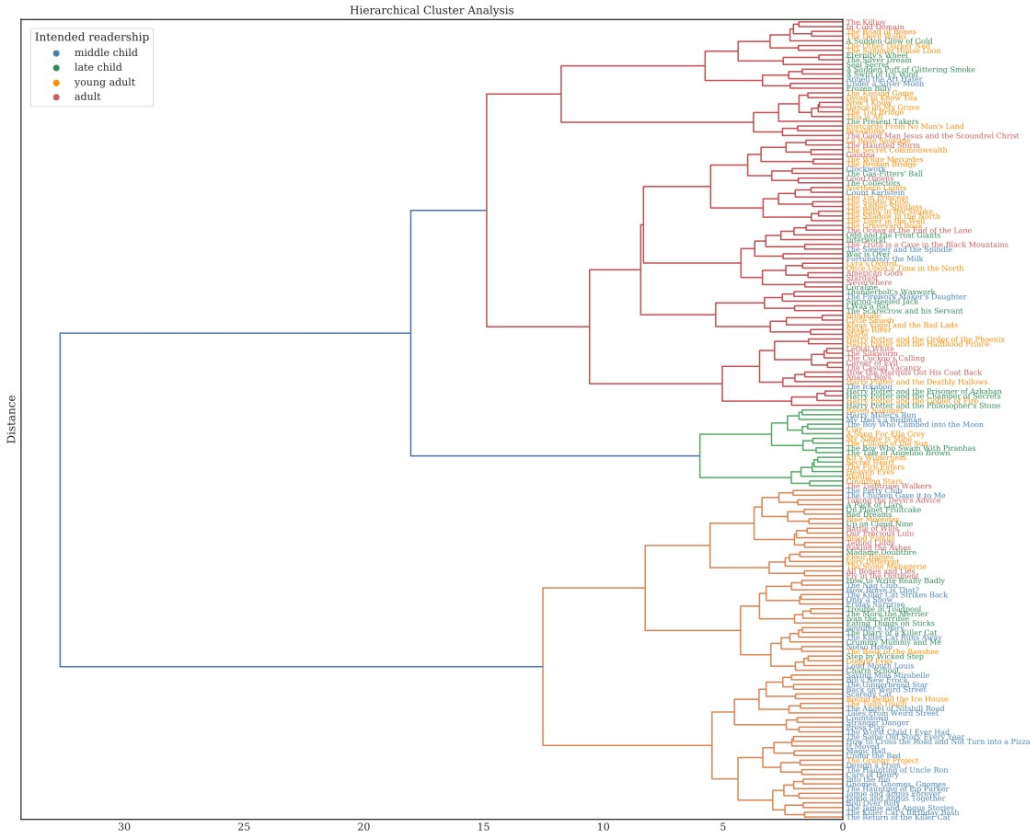


Figure 2.7: Hierarchical cluster analysis for books by English authors included in the corpus.

The second 'middle child' title that clusters together with books published for older readers is *The Ickabog*. Rowling had a specific readership in mind when developing this story, writing it chapter by chapter to read to her young children as a bedtime story, and she explicitly states on her website that the story is aimed at seven- to nine-year-olds. The *Harry Potter* books cluster closely with Rowling's books for adults when analysed together with other English authors on function words, but *The Ickabog* is not included in that cluster. In order to study this interesting case more closely, we will now zoom in on Rowling's oeuvre.

2.4.4 J.K. Rowling

Stylometry was thrust into the limelight in July 2013, after Juola used it to reveal that J.K. Rowling was the real person behind the British debutant Robert Galbraith. While he emphasises that stylometry does not provide a fool-proof method for author attribution, the result of his analyses led Rowling to admit she used the pseudonym to publish *The Cuckoo's Calling*. The analyses presented in Figure 2.6 confirm Juola's cautious conclusion: the writing style of *The Cuckoo's Calling* and its three sequels is very similar to Rowling's books for younger readers. However, the graph simultaneously illustrates Juola's caution: Rowling's books not only cluster among themselves but are also grouped closely together with books by Pullman and Gaiman. Rowling's most recent children's book, *The Ickabog*, is even separated from her other work. This analysis investigates the hypothesis that the stylistic difference between *The Ickabog* and Rowling's other books is influenced by more than just the age of the intended reader. We wonder whether Rowling's work shows a correlation between her writing style and the year of publication. Interesting to note here is that Rowling allegedly worked on the first draft of *The Ickabog* while writing the *Harry Potter* series. When the draft was completed, Rowling decided to step away from children's books and stored the manuscript in her attic. In the spring of 2020, she returned to *The Ickabog*, editing, finishing and publishing it in instalments over several months (Rowling, 'J.K. Rowling Introduces').

To further investigate the outlier of this work and to study where it is located stylistically when taking the publication date into account, we turn to a stylometric analysis of Rowling's oeuvre. Like Vandermeeren's and Kuijer's, Rowling's books show a great diversity in length and thus were divided into segments of 10,000 tokens. The analysis was conducted using 100 MFW. Figure 2.8 shows a clear divide between Rowling's *Harry Potter* books on the left and the *Cormoran Strike* series together with *The Casual Vacancy*, all books written for an adult audience, on the right side of the scatter plot. When we consider the books for young readers in more detail, most of the segments belonging to books for 'late child' readers (i.e. the first three *Harry Potter* books) gravitate more to the left of the graph when compared to most segments of the YA books (the last four). This is in line with the previous observation that the age of the intended reader of Rowling's books increases the more to the right the segments of that book are on the PCA – a trend that suggests that there is a correlation between style and the age of the intended reader. Interesting then is the location of the six segments of *The Ickabog*, which do not follow this trend. While this book is aimed at children younger than the readership of the first *Harry Potter* book (which Scholastic set from 9 to 12), in our PCA *The Ickabog* is located between segments of young adult and adult books.

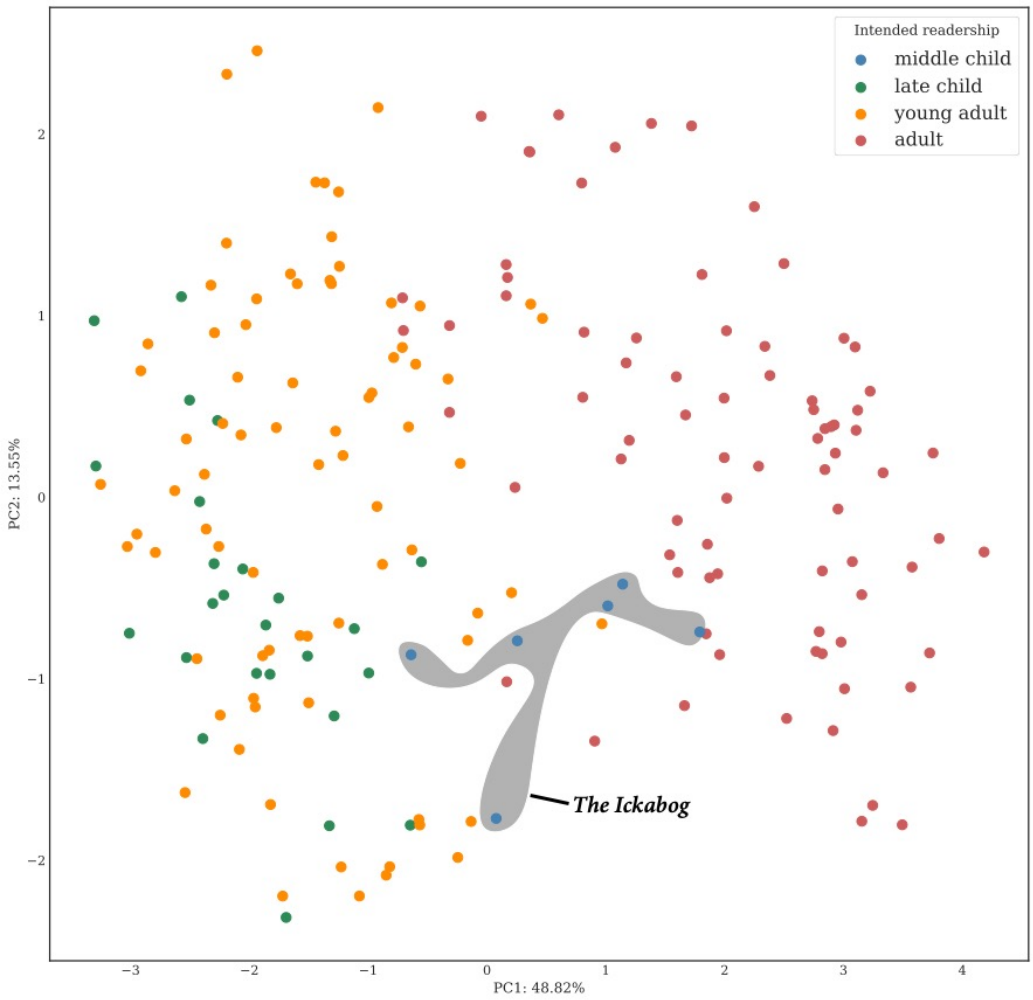


Figure 2.8: PCA for 10,000-token samples of books by Rowling (samples colour-coded according to intended readership). Note: PCA: principal components analysis.

Figure 2.9 shows the PCA for the segmented texts and it colour-codes the dots not according to intended age, but to publication date. First, we should note that the increase in the intended readership of Rowling's books runs largely parallel to their publication date. The *Harry Potter* series (intended for 'late child' and 'young adult') was published between 1997 and 2007, after which the adult books *The Casual Vacancy* and the *Cormoran Strike* series was published (2012–2018). This makes it difficult to pinpoint the most likely influence. From both points of view, *The Ickabog* appears as an anomaly in Rowling's oeuvre: it was published in the same decade as her adult books but is located closer to titles published around 15 years earlier. However, when we take into account the fact that this book was first written in that earlier time period, the PCA provides a more coherent picture. If we compare our two analyses (Figures 2.8 and 2.9) and if we consider the time when *The Ickabog* was written rather than its publication date, we find evidence that Rowling's evolving writing style runs more parallel to the period in which she was writing than to the audience she was addressing.

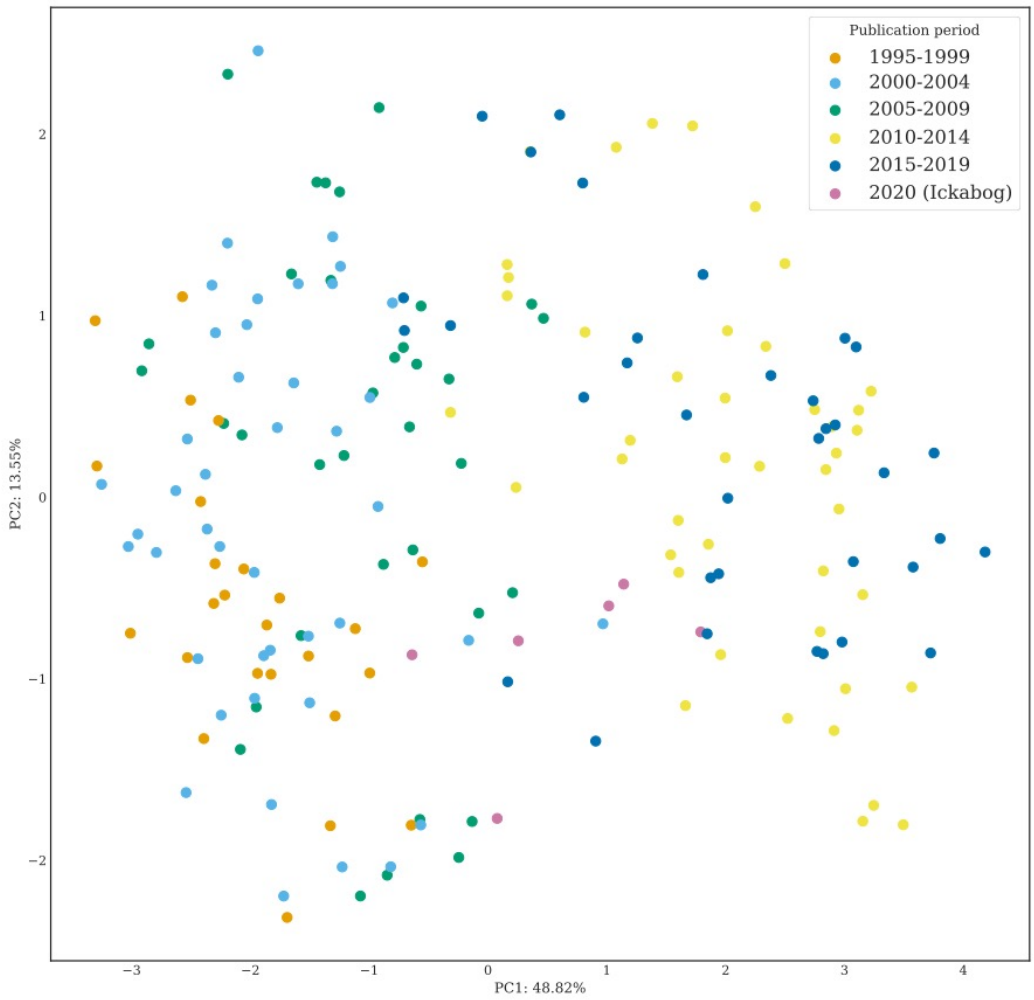


Figure 2.9: PCA for 10,000-token samples of books by Rowling (samples colour-coded according to publication period). Note: PCA: principal components analysis.

2.5 Conclusion

It should be emphasised that the stylometric techniques that we used are unsupervised. That is to say, no information was provided to the computer that could somehow influence the analyses. As a result, the computer has no knowledge of the author, the intended reader, or the publication date of any novel. Instead, the computer merely 'sees' the raw text. We have thus tried to establish from an unbiased perspective in what way the oeuvre of crosswriters can be stylistically typified. Both the analyses on the full English and Dutch corpora as the individual case studies show that the paratextual age of the intended reader is a discriminatory factor in the way that crosswriters' works are stylistically clustered. Based on the occurrence of function words, we see, moreover, that a correlation of the stylistic similarities with the age of the intended reader is usually stronger than that with the period or genre. This does not mean that those factors have no effect. In fact, when looking for explanations for outliers in the age-related clustering, we have turned to periodisation in particular (e.g., for explaining the unexpected positioning of *Count Karlstein* and *The Ickabog*). Moreover, we have seen that some authors (e.g., Vandermeeren and Rowling) turn to older readers over the course of time, so that rises in the age of the intended readership run parallel to the course of authors' writing careers. This complicates a straightforward investigation into the underlying influences for the clustering of these authors' novels.

Within the gradual age-related clustering that the stylometric analyses display, we can note further trends if we consider more fine-grained age differences. Young adult works tend to be stylistically more similar to adult works than to books for late and middle children. Books for middle children tend to cluster together quite distinctively. Our analyses belie the idea that it is only typical of children's literature to take audiences' needs into account or opt for a deliberately simple style. We argued that the choice for a simpler style may explain why some adult titles cluster with children's books for young readers and found evidence for this in paratextual and peritextual material (interviews, publisher's motto's). This simplification of style may be a conscious choice, motivated by certain publishers or series (e.g., *Wablieft* books, Bassington Stoke) or by an author's poetics (as was the case for Kuijer). Such stylistic choices are not limited to children's literature, but are also reflected in the diversity of adult literature.

Our analysis also raises new research questions and yields various opportunities for further research. Can the same trends be identified in a larger corpus of Dutch and English crosswriters, and do they also hold for older texts and books written in other languages? Moreover, several of the English texts have been translated into Dutch: do these translations cluster similarly according to age? Or can the impact of a translator influence the stylometric analysis? In this article, we focused on the discussion of results for analyses with 100 MFW, which comes down to distributions of function words. Raising this limit to 300 MFW did not produce significant differences. One may wonder if a consideration of particular content words would have an effect, for instance for genre, and if so, where the tilting point lies. In addition, aspects that we neglected in this article are potential effects of the author's gender, own age, region (Dutch books from Flanders vs the Netherlands) and publisher. These data are easy to collect and would allow for even more fine-grained analyses. A final important avenue that we see lies in authorship attribution and periodisation through stylometry. Children's literature, after all, has a rich history of pseudonymous and anonymous texts, as well as texts without publication dates. In any case, the field of children's literature studies still has a lot to gain from computers and researchers joining forces.

[i] While Anderson's dissertation was published in the 1980s, it was not until the early 1990s that these types of authors were referred to with the term 'crosswriters'. Sandra Beckett traces her use of the term back to a workshop led by U.C. Knoepflmacher at the Modern Language Convention of 1993 (Beckett *Transcending Boundaries*, xi-xii).

[ii] In order to check the reliability of our analyses, we also ran the analyses using both 50 MFW and 300 MFW. This higher number, however, does entail the risk that not only function words are selected, but also novel-specific content words. In order to minimise the risk of such words affecting our analyses, we manually deleted proper names and other novel-specific terms (e.g., 'Harry' and 'wizard', which are both highly frequent in the *Harry Potter* series). The results obtained from these analyses confirm the general trends that we were already able to observe for other analyses, in which we use the limited set of 100 MFW.

[iii] For our stylometric analyses, the frequencies of our manually edited list of 100 MFW were analysed across the different novels. The word frequencies were normalised according to the L2 norm and scaled with a TF-IDF vectoriser. Subsequently, the distances between novels were calculated using the Cosine similarity, a standardised and widely popular distance function, commonly used to calculate the stylistic resemblance between literary works (Jannidis et al.). The resulting frequency distributions were used as input for the algorithms constructing the HCA's and PCA's. Specifically for performing the hierarchical clustering procedure, we employed Ward's method. For a detailed overview of the modules and packages used, we refer the reader to our publicly available code (DOI: 10.5281/zenodo.5521241).

[iv] Our translation. Original text: "Ik kan mij perfect identificeren met een kind van 9 jaar".

[v] Full list of Dutch words taken into account for our stylometric analysis (words marked with '#' are excluded): aan, al, alleen, als, #ben, bij, daar, dacht, dan, dat, de, die, doen, door, een, eens, en, er, even, gaan, geen, ging, goed, #haar, had, heb, hebben, heeft, #hem, het, hier, #hij, hoe, #hoofd, #hun, iets, #ik, in, is, ja, #je, #jij, kan, keek, kon, kwam, maar, #man, #me, meer, met, #mijn, misschien, #moeder, moest, moet, naar, niet, nog, nooit, nu, of, #ogen, om, #ons, ook, op, over, stond, te, tegen, toch, toen, tot, #twee, #u, uit, #vader, van, voor, vroeg, waar, waren, was, wat, #we, weer, weet, #weg, wel, werd, wil, zag, #ze, #zegt, #zei, #zich, zijn, zo, zou.

[vi] Our translation. Original text: "Het heeft me relatief weinig moeite gekost om de plot te verzinnen en duidelijke taal te gebruiken. Belangrijk voor de doelgroep is het positieve gevoel dat nadien bovendrijft. In de zin van: 'Wow, ik heb een boek gelezen!'"

[vii] Our translation. Original text: "Het dochtertje van de wasvrouw is een poging om [...] eenvoudiger te gaan schrijven. [H]et staat natuurlijk interessanter als je ingewikkelder, komplekser, moeilijker schrijft [...]. [M]aar als je voor kinderen gaat schrijven, dan kun je dat niet maken, die prikken daar gewoon doorheen".

[viii] Studying both Dutch and English authors in the same analysis requires a time-consuming pre-processing, the method of which lies beyond the scope of this article (see Cinkova and Rybicki).

[ix] There is also a large variety in text length which would give a misrepresentation of the size and significance of any resulting clusters.

Chapter 3 Putting the Sorting Hat on J.K. Rowling's Reader

A Digital Inquiry into the Age of the Implied Readership of the Series[2]

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.

Context – This chapter expands on chapter two in exploring which digital methods can be used to look for differences in texts for readers of various ages by studying both the formal features and content of a small collection of unannotated texts (RQ 2.2 and 2.3). As pointed out in the introduction, chapter three uses the term 'implied reader' rather than 'intended reader' to refer to the readership as it is constructed by a text. Although this chapter examines the data using multiple digital methods, the cohesion of these analyses may not be as strong as it could have been. The argument made in this chapter would have benefited either from a closer collaboration on both analyses or from a more detailed discussion in two separate publications. The latter option would also permit more scope for supplementing the quantitative analyses with close reading. This is promised in the introduction but was only applied to the semantic analysis. Furthermore, the method of traditional topic modelling, which is not used in any of the other chapters in this dissertation, could have benefitted from a more extensive discussion to be consolidated as a tool to study the influence of the age of the intended reader of a text. To gain a greater understanding of the construction of age *by* a work of fiction (RQ 1.1), the analyses are limited to one author's oeuvre and thus eliminate paratextual features that might affect the style and topics of the texts in question, such as individual authors, publisher, or genre. Besides answering the research questions stated above, this chapter also contributes to defining the

[2] This chapter is based on: Haverals, Wouter and Lindsey Geybels. 'Putting the Sorting Hat on J.K. Rowling's Reader: A Digital Inquiry into the Age of the Implied Readership of the Series'. *Journal of Cultural Analytics*, vol. 5, 2021, pp. 1–30.

vague terms of ‘complexity’ and ‘simplicity’ that are often used in the discussion of differences between fiction for children and adults.

3.1 Introduction

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). 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" (in Nel, 51). *Harry Potter* was thus conceptualised as crossover literature (Beckett *Crossover Fiction*) 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 recognised an evolution in complexity in the *Harry Potter* series, which they relate to the increasing age of its readers.[i] 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.

3.2 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" (20),[ii] while Seth Lerer defines it as literature that is "read by children" (11). 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" (2014). 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" (ibid.). 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" (Prince, 404). A second interpretation of the implied reader coined by Schmid is the "ideal recipient who understands the work" (2014). 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'[iii]. 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 (*Crossover Fiction*) emphasises the power publishers have in determining the implied audience of children's literature. 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 (ibid., 196). 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 (Eccleshare, 7–8). 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" (Fraser). 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' (Errington, 46). 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 categorise 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 (Common Sense Media, 'How We Rate'). 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" (Common Sense Media, 'Harry Potter'). 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" (ibid.). 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.[iv] These values correspond to grade levels and can therefore be converted to age ranges. Figure 3.1 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.[v] 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" (Errington, 207). Scholastic further provides a categorisation of the series according to grade level on their website, which provides teaching tools (Scholastic). Second, the Central File of Children's Books (Centraal Bestand Kinderboeken – CBK) 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 3.1 as well.

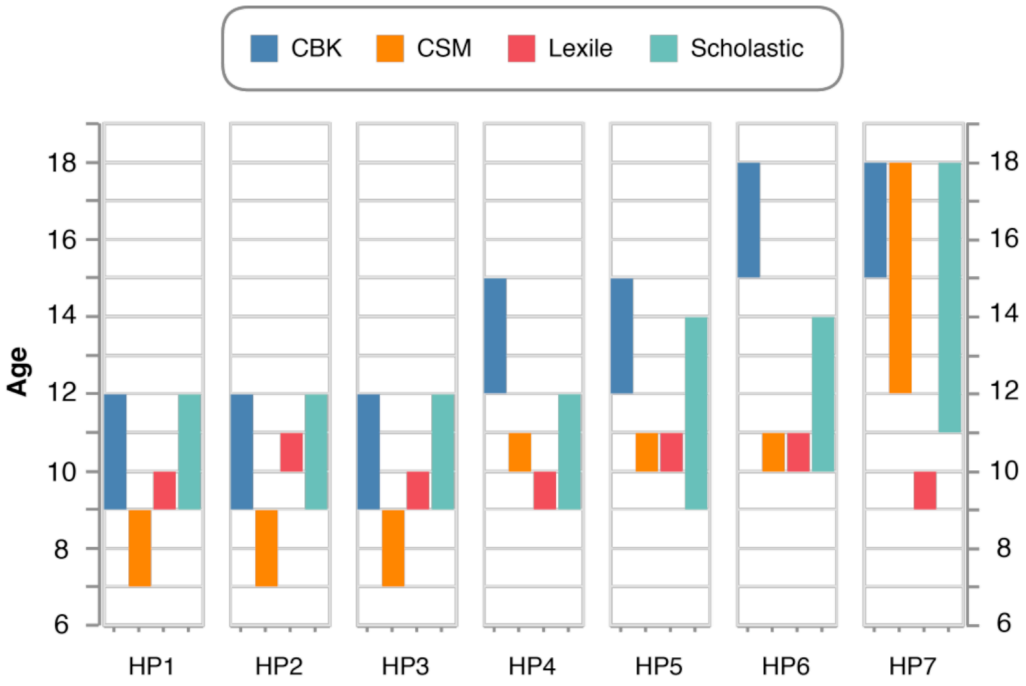


Figure 3.1: Age of the implied reader of the Harry Potter series as recorded by Centraal Bestand Kinderboeken (CBK), Common Sense Media (CMS), Lexile, and the series' American publishing house, Scholastic.

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 3.1 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 *Children's Literature*, De Vriend, Fastenau). Stephen Krashen even considers the Lexile measure "unnecessary and potentially harmful." 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.

3.3 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, Fry, Yi Ma and Loftus). 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?[vi] The selection of available digital analyses we made is informed by a paper by Wanner et al. in which the writers present and evaluate a tool for assessing age suitability of books. 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 (Bailin and Grafstein). 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.

3.4 Form in the series

3.4.1 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 (5). 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).[vii] 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 3.2) 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[viii] 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.

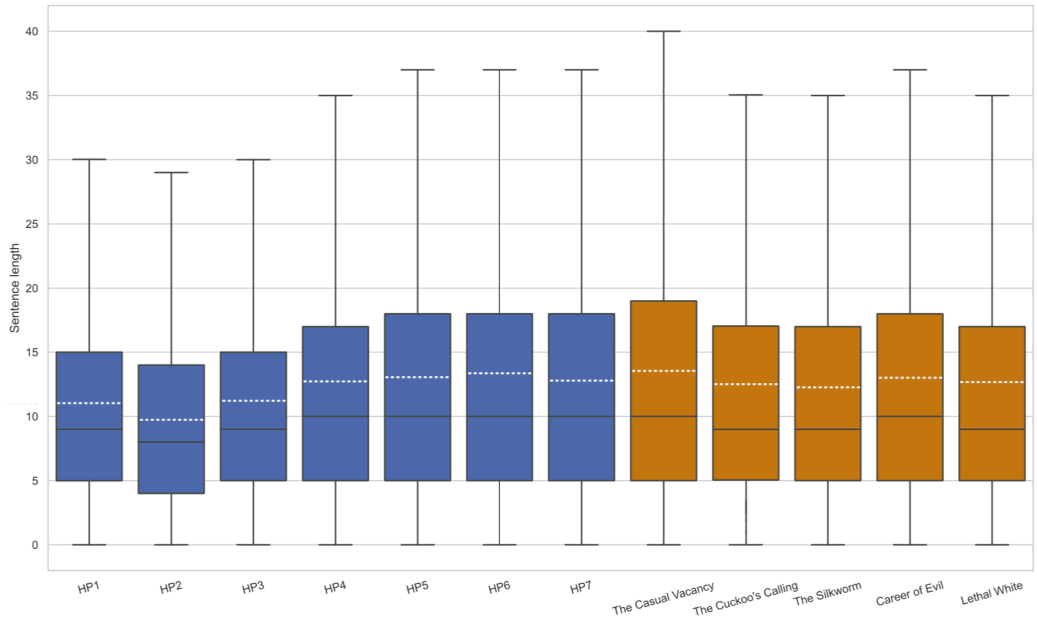


Figure 3.2: Sentence length of each book in the Harry Potter series and J.K. Rowling's adult fiction (ordered by publication date). Average sentence length is represented by a white dotted line.

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 (Bailin and Grafstein, 13). 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 (65–80). 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 3.3 shows for each book the proportion of sentences with at least one and more than two subordinate clauses.[ix] 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).

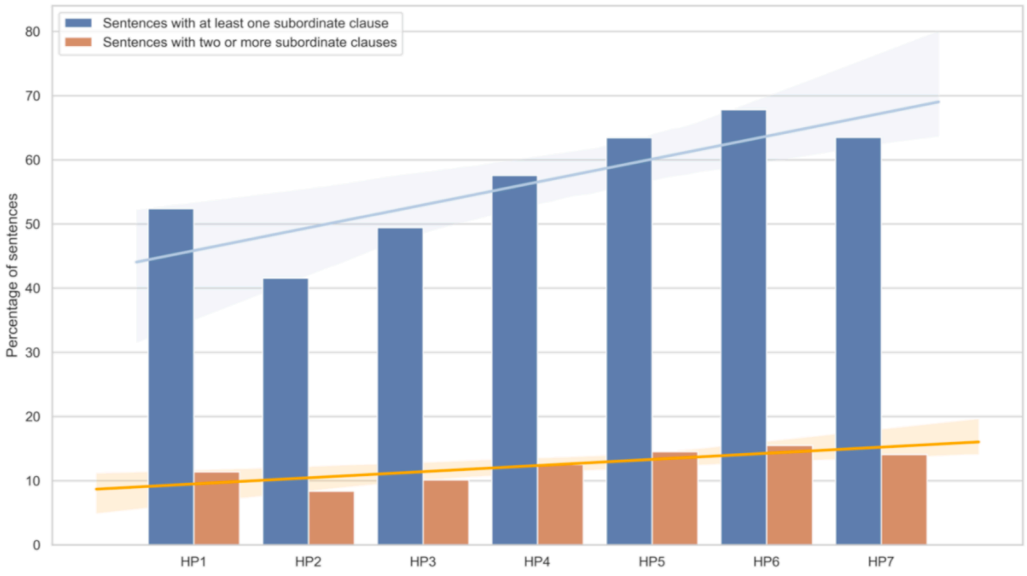


Figure 3.3: Ratio of sentences per novel that contains subordinate clauses. Syntactic structure was analysed using the Berkley Neural Parser *(Kitaev and Klein 2018).*

3.4.2 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. 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 (see Wendell 'Studies', Templin). 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 and Jarvis). 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.[x] Ideally suited for this purpose is the Moving Average Type-Token Ratio (MATTR), developed by Covington and McFall. 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.[xi] Figure 3.4 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 3.4, 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.

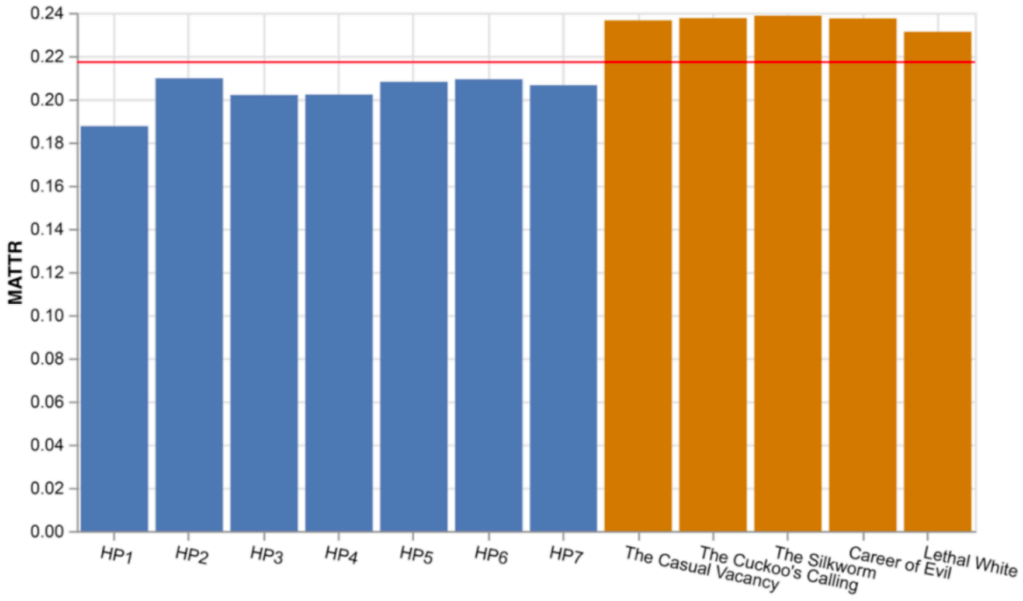


Figure 3.4: MATTR scores for the individual books in the Harry Potter *series and Rowling's adult novels. Calculated using a 10,000-word window size. The novels appearing on the x-axis are ordered chronologically. The red line shows the mean (.22).*

3.4.3 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. include the calculation of the Automated Readability Index (ARI). This readability test was developed by R.J. Senter and Edgar A. Smith in 1967 and combines information on the number of characters, words and sentences in a text.[xii] 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 (Fry). Librarians use them to aid visitors in their search for suitable reading materials. Crossley et al. 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 their discussion of readability formulas, they refer to a study they conducted in 2017 in which they demonstrated the benefit of readability formulas that use "features that measure lexical and syntactic constructs, text cohesion, sentiment, topic analysis and semantics" (3).

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 (Gunning), Dale-Chall (Dale and Chall 'Formula', *Readability*), the Simple Measure of Gobbledygook (McLaughlin), the Coleman-Liau Index (Coleman and Liau) and the Flesch-Kincaid formula (Kincaid et al.). These formulas output a value that corresponds to the reading abilities of a student within the U.S. grade level system. Table 3.1 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 3.1: 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 criticised.[xiii] 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 3.5 shows the scores obtained for the above-mentioned readability formulas directly translated to the U.S. grade scale.[xiv] 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 τ =.39, p<.001). Flesch-Kincaid (τ=.31), ARI (τ=.28) and Coleman-Liau (τ=.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 recognised 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) (Mesmer). 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.

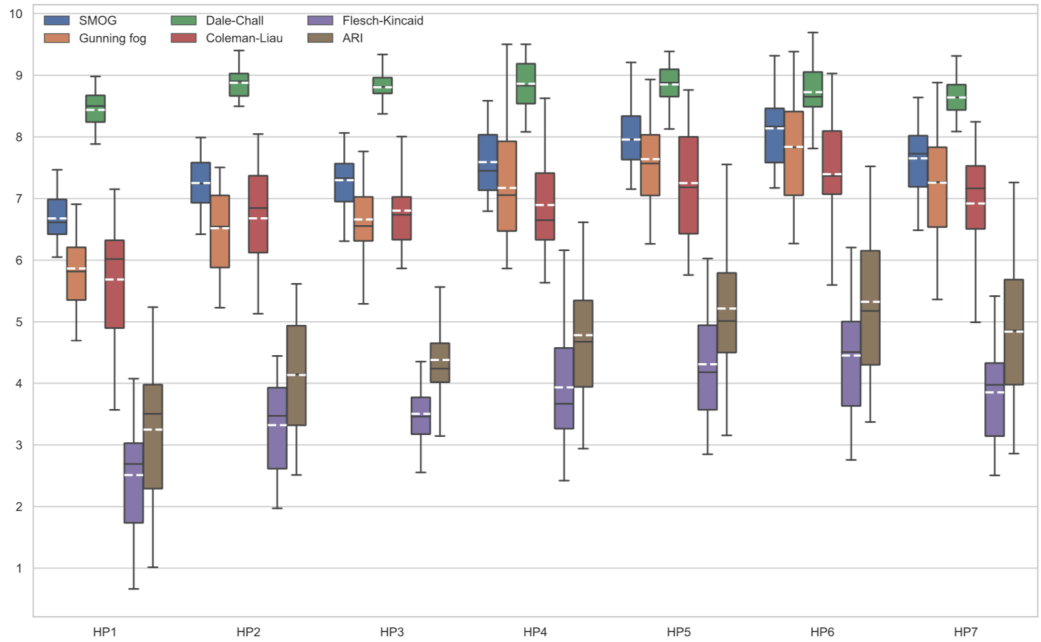


Figure 3.5: Readability formulas applied to 25 200-sentence samples taken from the individual books in the Harry Potter *series.*

From the averages in Figure 3.5 (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 3.4). 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 (Duffy). 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 (Mesmer).

Similar discrepancies can also be identified if we look at the age attributions set by recognised institutions. Table 3.2 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 3.5, 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		10-11		9-10
ARI	9-10		10-11		11-12		10-11
Coleman-Liau	11-12		12-13				
Gunning fog	11-12		12-13		13-14		
SMOG	12-13		13-14			14-15	13-14
Dale-Chall	13-14	14-15					
All*	10-11	11-12	12-13				

Table 3.2: 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.*

3.5 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.[xv] The analysis of topics is one of the components Crossley et al. 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 (Zhu et al.),[xvi] to identify word clusters.[xvii] 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.[xviii] 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 3.3 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.

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 connects non-human characters, including animals, to childhood (*Rhetoric*). 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 3.6 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, the evolution of the topic of magical details, such as magical creatures and animals, is closely linked to the age of the implied reader (117). 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 3.6 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' 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. 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.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Decrease (100 topics)	dog dogs puppy cat **animals**	**game** team ball play players	class **school** teacher students classroom	doctor **hospital** nurse patient patients	kitchen bathroom bedroom shower stairs	shit gon ass whispered babe
Decrease (200 topics)	letter **letters** envelope read paper	**school** year college summer grade	class teacher students **classroom** classes	cat cats **animals** animal shelter	**game** team ball play players	dog puppy pet **animal** tail
Decrease (300 topics)	**dog** pet tail barking fur	**food** eat plate eating meal	class classes **classroom** lunch today	**game** ball play players baseball	**school** football grade teachers college	**candy** chocolate tickets cookies cookie
Increase (100 topics)	**wedding** dress marry married bride	soldiers soldier **military** tent riffle	tears loved whispered cry crying	replied answered exclaimed explained shouted	wife husband daughter married **marriage**	church altar soul **souls** angels
Increase (200 topics)	**death** died die killed alive	army **war** battle enemy troops	truth **conversation** feelings tone trust	sword blade **swords** hilt dagger	daughter husband daughters birth age	mage ivory silver pack healing
Increase (300 topics)	**ring** finger diamond engagement rings	**death** died funeral die grave	answered stated explained group responded	tones gentleman demanded **fear** features	nature **social** order course knowledge	daughter daughters birth age parent

Table 3.3: 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.*

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, the school setting in the *Harry Potter* books emphasises the power structure of adult over child (*Power*). The decrease of this topic as the series progresses indicates a moving away from childhood and towards adulthood. Figure 3.6 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 *Rhetoric*) 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 3.6: 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 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 externalised and overcome. 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, 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 'Harry Potter', Cockrell, and Behr). Figure 3.7 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 and Lee Talley 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 (*Rhetoric*) draws between sexuality in general fiction and food in fiction for children (42) is not visible in the topic model. While topics on food have already been shown to decrease, the topics included in Table 3.3 do not support a complementary increase in sexuality with regards to the *Harry Potter* series.

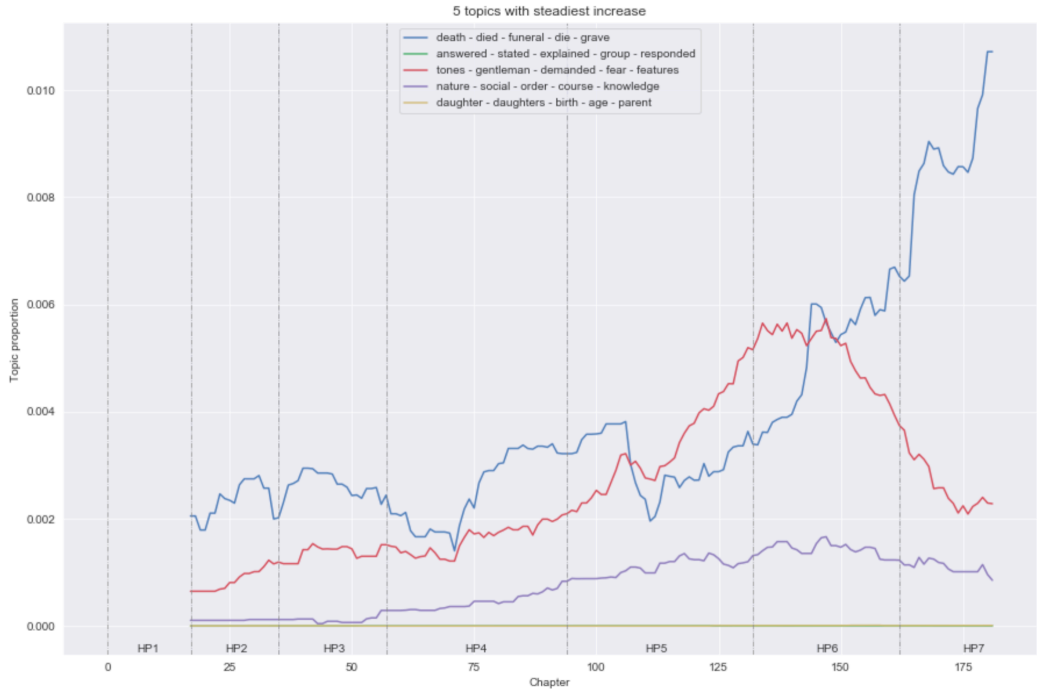


Figure 3.7: Evolution of the five topics with the steadiest increase in the Harry Potter *series; only the top three show a meaningful evolution.*

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" (100). 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 (Beckett *Crossover Fiction*, 67). 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.

3.6 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. Rebecca Butler 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 (67).

[i] The exact nature of this relation has been subject to a debate that lies beyond the scope of this article. Kümmerling-Meibauer 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 (*Crossover Fiction*) and Victor Watson 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.

[ii] 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.

[iii] 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.

[iv] 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, 4).

[v] 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.

[vi] 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/

[vii] For this calculation, the English-dedicated sentence and word tokeniser 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 tokenisation at the word-level, all non-alphanumeric characters were discarded.

[viii] 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.

[ix] The syntactic structure analysis was performed using the *Berkley Neural Parser* (Kitaev and Klein), 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.

[x] Minimising 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 (McCarthy and Jarvis, and Kyle 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.

[xi] 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 and McFall (2010), 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 (97). The calculation of MATTR was performed, using 'lexical-diversity', a package for Python developed by Kristopher Kyle (https://github.com/kristopherkyle/lexical_diversity).

[xii] 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.

[xiii] Reading speed is one of the components Crossley et al. include in their study of alternatives to classic readability measures. See i.a. Bertram et al. and Schriver.

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

[xv] 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*.

[xvi] 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.

[xvii] 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.). 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.

[xviii] 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.

Chapter 4 Determining Author or Reader

A Statistical Analysis of Textual Features in Children's and Adult Literature[3]

Abstract – Due to the nature of literary texts as being composed of words rather than numbers, they are not an obvious choice to serve as data for statistical analyses. However, with the help of computational techniques, words can be converted to numerical data and certain parts of a text can be examined on a large scale. Textual elements such as sentence length, word length and lexical diversity, which are associated by scholars on the one hand with the writing style of an individual author and on the other with the complexity of a text and the intended age of its readers, can thus be subjected to statistical evaluation. In this paper, data from little under 700 English and Dutch books written for different ages is analysed using a statistical linear mixed model. The results show that the textual elements studied are better qualified to detect the age of the intended reader of a text than the identity or age of the author.

Context – The statistical analyses in this chapter substantiate the results from the analyses of chapters two and three, in which a stylometric analysis of part of the corpus and a formal analysis of the oeuvre of one author indicated that there is a link between the intended age of a text and its formal features. The number of texts used as input for these analyses is the largest of the chapters in this dissertation, consisting of almost all texts from the corpus. Using the unannotated text files as input for linear mixed models, I looked for a statistically significant correlation between the age of the intended reader and a set of formal features which are informed by chapters three and five (RQ 1.1 and RQ 2.2). While the argument for using linear mixed models was somewhat superfluous for the initial audience of this chapter, this addition ensures that the analysis is also accessible to children's literature scholars, who are often less familiar with statistical methods. The analyses in this chapter stay within the confines of distant reading given a discussion of outliers similar to chapter two would move the discussion beyond the constraints for the original publication.

[3] This chapter is based on: Geybels, Lindsey. *Determining Author or Reader: A Statistical Analysis of Textual Features in Children's and Adult Literature*. CEUR Workshop Proceedings of the Computational Humanities Research Conference, 12–14 Dec. 2022.

4.1 Introduction

Statistical tests are traditionally associated with exact science or sociology; for example to ascertain the effect of a treatment on the growth of a certain plant or to determine the efficacy of a school-based smoking prevention curriculum. Over the past few decades, the applicability of statistics in making quantitative decisions has expanded through innovations in technological, and specifically computational, areas. While its subject matter contains a myriad of data that can be used as input for computational analysis, literature is one of the fields not traditionally studied through mathematical calculations. However, the use of computational techniques allows literary researchers to "change slippery words into more absolute numerical […] substitutes" (Potter, 415). Possibly the most popular implementation of this method is found in the field of stylometry, which occupies itself with the study of linguistic style. By applying statistical analysis to specific features of a set of texts, stylometry is often employed to attribute authorship, either to anonymous or disputed documents. Due to the nature of this research, which often studies a handful of texts by as many authors to determine their stylistic proximity to an anonymous text, the analyses tend to be small-scaled (Levitsky and Melnyk) with a focus on authorship attribution of general literature.

This paper fits into the CAFYR (*Constructing Age For Young Readers*) research project and borrows from it the extensive corpus of 692 titles of children's, young adult, and adult literature written by Dutch, English and Flemish authors. To avoid the traditional treatment of the study of children's literature as being isolated from literature for adults, the corpus is focused on cross-writers; authors who write for children of different ages as well as an adult readership. Both the focus on these authors and the number of texts exceed the scope of many previous studies. For each title in the corpus, the average sentence length, average word length, measure for lexical diversity and ratio of dialogue versus narration is extracted and used as input for a linear mixed model. Linguistic features have been used in stylometric analysis to study several pieces of metadata. First and foremost, while the analysis of word frequencies, most often in the form of function words or n-grams, is generally considered to be more reliable in the practice of authorship attribution (Bozkurt et al., Grieve), previous research has proposed the use of simple token-based lexical features and vocabulary richness as markers to quantify an author's writing style (see Kjetssa, Forsyth et al., Chaski). Second, questions of chronology within the oeuvre of one author, corresponding to the changes in the writing style of an author as they age, have been answered by so-called stylochronometry (van Hulle and Kestemont). Finally, the same features are also often used to determine the readability of a text, a practice that is more common in children's literature and is connected to the age of the intended readership. In a stylometric analysis of the oeuvres of ten British, Dutch and Flemish authors, Haverals et al. found that the style of texts often correlates with the age of the intended reader. To judge if the textual features listed above, used both for authorship attribution and text readability, are better at predicting the author, their age or that of the reader, this paper reports on statistical analyses using linear mixed models while also investigating a possible difference in language. The question this paper tries to answer, in other words, is whether sentence length, word length, lexical diversity and the ratio between dialogue and narration are better suited to determine issues of authorship or of readability.

4.2 Previous research

Previous research using statistical techniques to analyse children's literature is scarce but not absent. Roger Clark identified a striking level of interest of feminist social science in a quantitative approach to children's literature (Clark). His literature review contains a list with over thirty articles featuring research into sexism, sex-roles, gender and other feminist social topics by 'counting', as Clark puts it somewhat simplistically (see also Hillman, Diekman and Murnen). While not all of these studies include statistical tests in their methodology, a number of them do and the field of social sciences seems to be in a leading position when it comes to statistical analyses of children's literature (Crisp et al.). In general literature, quantitative analyses of textual features have mainly been employed to study authorial style. In the field of computational stylometry, authorship attribution has been a hot topic for several decades, with different researchers defending the validity of certain textual features, while discarding others, in the search of a text's true author. Because anonymous or pseudonymous authors are rarely found in children's literature, it has yet to be considered as fruitful data for quantitative studies of textual features. However, these analyses can reveal much more about a text than merely the author's subconscious writing style. Haverals et al. used a stylometric analysis on the oeuvres of ten authors who write English or Dutch books intended for different readerships, ranging from 'middle child' (children aged 6 to 8) to 'adult' (people aged 18 and up). They looked not only for clustering according to the individual author, but included the age of the author and the age of the intended reader as features to their analyses. From several case-based analyses as well as from the corpus-wide analysis they performed on English titles, they concluded that stylometric analysis can be fruitful when investigating issues of, what I will call, readership attribution; the task of identifying the age of the intended readership of a text.

Most often, stylometric analyses are based on frequencies of function words, as is the case for Haverals et al.'s study. However, in children's literature, textual features including sentence length, word length and lexical diversity are more commonly used to determine the writing style or complexity of a text. These features are often consciously manipulated by the author or editor of a book in connection with literacy and education. Guidelines for the readability of a text, or readability formulas such as ARI, Gunning fog and Dale-Chall in large part rely on these values to categorise texts according to an appropriate reading age. Scholarly interest in this topic has produced mainly small-scale studies which do not rely on statistical analyses (see Meyer, Fry, Yi Ma and Lofthus). A notable exception is Celia Catlett Anderson's dissertation on style in children's literature. She performs a statistical analysis of several textual features from both books written for children and for adults, including sentence and word length, lexical repetition and the amount of dialogue, corresponding to the features studied in this paper. This paper aims to expand on previous research by using a larger dataset of textual features from 692 texts. Furthermore, whereas most studies into writing style have focused on a single language, the analyses below look for any significant differences between Dutch and English texts, based on sentence length, word length, lexical diversity and the ratio between dialogue and narration.

The main idea underlying studies into authorship attribution in the field of stylometry is that "authors have an unconscious aspect to their writing style" (Holmes 'Evolution', 111), certain features that they include in their writing without being able to actively manipulate them. Even before the emergence of computational methods to aid in the quantitative study of texts, the analysis of simple token-based lexical features, including sentence length and word length, was

being used to attribute authorship (see Mendenhall, Yule). Although the length of lexical units is generally not deemed to be a reliable indicator to determine a text's authorship, there have been studies that nuance this view (Mannion and Dixon). Furthermore, the distribution of word length was one of the elements used by Patrick Juola in 2013, when he brought the study of authorship attribution to the attention of people outside of literature and linguist departments. When it comes to the study of text complexity in children's literature and, closely connected to it, the age of the intended reader of a text, sentence length is a more commonly used measure. According to Colleen Lennon and Hal Burdick, "the best predictor of the difficulty of a sentence is its length" (5). In a study of children's book publishers, Anderson relates the average length of sentences in a book to the age of its readership. Through interviews with several publishers, she found one of the most common requests to children's books' authors is to shorten sentences to make the text simpler and more accessible to young readers. Guidelines like these determine the reading level of books. However, as each child develops reading skills at its own pace, this does not translate directly to the age of the reader but there is a close connection between both. This suggests that the average sentence length of a text is often, contrary to the assumption of early studies in authorship attribution, under conscious control of the author.

Lexical diversity, as a measurement of the number of different words in a text, is most simply represented by the ratio of the number of unique words (types) to the total number of words (tokens) (Johnson). Several more complex formulae exist to measure lexical diversity, which in addition consider the number of words that occur a specific number of times, their frequencies and arbitrary constants (Grieve). Parallel to the lexical features of sentence and word length, the vocabulary richness of a text has been disputed as well as confirmed as a reliable indicator in the search of a text's author (see Chaski, Gregori-Signes and Clavel-Arroitia). In addition to the standard authorship attribution problem, David Holmes identified a second use for stylometric research: chronological problems, or the hypothesis that stylistic features develop during an author's life ('Evolution'). It was this that Tallentire addressed in a study based on the ratio between hapaxes and tokens to measure lexical diversity, where he concluded that lexical diversity decreases with age of the author ('Confirming Intuitions'). Victoria Johansson confirms this correlation in her study conducted on the writing and speech of children aged 10, 13, 17 and adults. However, from a developmental point of view, she found that lexical diversity increases with age, attesting to vocabulary development in children. Lexical diversity is not only linked to the creation of a text but also its consumption and the pedagogical role of children's literature. In Wanner et al.'s 'Age Suitability Analysis', several aspects that play a part in determining the appropriate reading age of a book are listed. One of them is linguistic complexity; the "difficulty of the writing style" which is measured using a variation on type/token ratio (Wanner et al.). Texts intended for younger readers contain more repetitions and thus will present a more limited vocabulary (Bland).

The ratio between dialogue and narration is not traditionally connected to studies of authorship or text complexity, but according to Rita Ghesquière, this measure is influenced by the age of the intended reader, as she has found that books for younger readers usually contain more dialogue when compared to books aimed at adults (Ghesquière).

4.3 Collecting data

The analyses in this paper are conducted on 692 works of prose fiction published by Dutch, British and Flemish authors between 1970 and 2020. To investigate the relation of metadata concerning author and reader to textual features, the following information is stored for each book: language, year of publication, the age of the author at the time of publication and the age of the intended reader. Due to a movement in the United Kingdom which opposes children's literature publishers' strategy of 'banding' their books by adding explicit age guidelines to the cover, the latter was found on the physical books for only a small percentage of the texts, which were almost exclusively Dutch books. For the remaining texts, the age of the intended reader was retrieved from other sources, if available, in order of precedence: publisher's catalogues, author's websites, the Dutch database for children's books (CBK) or websites of booksellers.

After collecting metadata on the extensive corpus, several features were extracted from the individual texts, which were first stripped of paratext and chapter headings: sentence length, word length, lexical diversity, and the ratio between direct and indirect speech. For the first two, tokenisers from the Natural Language Toolkit were used. To calculate lexical diversity, the Moving Average Type-Token Ratio (MATTR) was used (Covington and McFall). By computing the average of the type/token ratio of a moving window with fixed length, this method resolves the fallacy of many other formulae which do not consider the length of the text being analysed. The text was first lemmatised using *spaCy*, which is sensitive to different parts of speech. Covington recommends a small window size when analysing patterns of repetition, or a large window when trying to determine the size of the writer's vocabulary (Covington). In the analyses for this article, the window size is set to 1000 words, a value close to the word count of the shortest book in the corpus, *Sinclair Wonder Bear* by Malorie Blackman (1036 words). The ratio between direct and indirect speech was added to the present study as a control feature as it has not been considered in authorship attribution studies, and thus no correlation with the author is assumed. The distinction between narration and dialogue in the texts in the corpus was made by means of quotation marks. Unfortunately, a large part of the texts was digitised from scans and small punctuation marks are subject to errors due to the OCR (optical character recognition) process. To minimise this error as much as possible, manually annotated material was used as input where available. About one third of the corpus is annotated to study aspects of characterisation in the broader research project. These annotations include the identification of characters' speech versus narration and can thus be used to extract the number of words in either category.

4.4 Method: why mixed models

Basic statistical models, such as linear regression, are suitable to analyse simplistic and clean data, which is a rarity in life sciences. Much more often, data is non-independent, which arises from repeated measures or a hierarchical structure; a study can rely on multiple measurements of the same subjects at different times or data points can be grouped. Ignoring non-independence by using standard linear regression in these cases means that not all the variation in the data will be captured. Linear mixed models (LMM) provide an answer to this problem by considering random effects as well as the fixed effects recognised by linear statistical models and compiling all individual results into one model. LMM distinguish within-group variability from between-group variability, capturing the total variability of the dataset, and are thus well-suited for analysing the corpus of this paper. With a total of 692 titles written by 27 authors in two

languages, the textual features that are extracted from the texts must be considered in a hierarchical structure. While linear regression could be used to answer questions such as: 'Do authors construct longer sentences according to their own ageing process?', this would only hold for the analysis of one single author. Chucking the data from all 27 authors together while not accounting for between group variability means that we ignore an author's individual style and linguistic features inherent to different languages.

In addition to the advantages of LMM when working with non-independent data, the model is noted for its robustness when dealing with missing values. Despite the number of available sources reporting on the age of the intended reader as detailed above, this information proved to be unobtainable for several books in the corpus. The sixty titles in question, of which only three written by Dutch or Flemish authors, are all assumed to be targeted towards children or adolescents due their imprint exclusively publishing children's literature. However, the setup of this study is not to look at the dichotomy of literature for adults versus literature for children, but rather investigate a further categorisation of fiction for young readers into narrower age ranges. For this reason, no value is recorded for these sixty books. Common practice dictates that if linear regression was to be used to analyse the dataset, titles for which the age of the intended reader is unknown would either be removed from the model or their missing value would be estimated based on other entries (Molenberghs et al.). However, when working with LMM, these titles are not skipped nor is the age of their intended readers estimated.[i]

4.5 Results

4.5.1 Sentence length

In a basic analysis comparing the sentence length of Dutch and English texts, a t-test shows language to have a significant effect ($p < 0.0001$). Overall, the average sentence in the English part of the corpus counts 9.051 words, while the mean in Dutch is 8.072 words per sentence. In a LMM fitted with author as a random effect and average sentence length as outcome, the fixed effects of language and the age of the reader, as well as the interaction of both features, are highly significant ($p < 0.0001$). The age of the author proves to be insignificant ($p = 0.056$) and is thus excluded from the model. In the resulting model, the intra-cluster correlation is 0.463 (between author $\sigma = 1.192$; within author $\sigma = 1.283$). When splitting the model by language, the age of the author has a significant effect on the average sentence length of the Dutch part of the corpus ($p = 0.002$).

4.5.2 Word length

When performing a naive t-test on the average word length, there is a significant difference between Dutch and English texts ($p < 0.0001$). Overall, English words are 4.018 characters long while Dutch words are longer, 4.306 characters. However, the LMM with author as random effect and average word length as outcome does not show a significance for the main effect of language ($p = 0.600$). The age of the author has a significant effect on the average word length ($p < 0.0001$), regardless of language. In terms of average word length, the intra-cluster correlation is 0.35 (between author $\sigma = 0.093$; within author $\sigma = 0.127$). On the other hand, the age of the intended reader and its interaction with language are highly significant ($p < 0.0001$). When splitting the model by language, the effect of the age of the author on the average word length becomes less

significant (p = 0.021) for Dutch texts while the age of the intended reader becomes slightly less significant (p = 0.007) for English texts.

4.5.3 Lexical diversity

A Welch Two Sample t-test indicates that language has a significant effect on the lexical diversity of texts (p < 0.0001) with an average of 0.288 for English and 0.304 for Dutch texts. In a LMM, the age of the author has no significant effect on lexical diversity (p = 0.061) and is excluded from the model. Once again, the resulting model shows a highly significant effect of the age of the intended reader (p < 0.0001) and a significant interaction between this feature and language (p = 0.011). Language by itself is not a significant effect (p = 0.964). The intra-cluster correlation is 0.500 (between author σ = 0.024; within author σ = 0.023).

4.5.4 Ratio dialogue vs. narration

Once again, a simple t-test suggests a significant effect of language on the ratio between dialogue and narration. With an average of 0.379, English texts have a higher average value than Dutch texts (0.310). However, none of the fixed effects of language (p = 0.190), age of the intended reader (p = 0.316) and age of the author (p = 0.250) have a significant effect on the ratio between dialogue and narration in a LMM. The correlation between titles of the same author considering this feature is low (0.152).

4.6 Discussion

A statistical analysis of sentence and word length confirms the hypothesis that the age of the reader has a larger influence on writing style than the individual author or their own age. The correlation between titles of the same author is moderate when considering average sentence length and the age of the author proves to only have a significant effect on the Dutch part of the corpus. Similarly for average word length, the correlation between titles of the same author is low. This supports Jack Grieve's conclusion of his evaluation of authorship attribution techniques; namely that "the value of a single measurement of average word- or sentence-length, [...] appear[s] to be of little use to investigators of authorship" (259). In contrast to the findings discussed above, the effect of the age of the author on average word length is smaller for Dutch when compared to the English texts. A more significant effect on the average sentence and word length is the age of the intended reader, both isolated and in the interaction with language. This means that the effect of one feature depends on the other. Overall, the LMM estimates that words and sentences lengthen as its readership ages, confirming the hypothesis that these features are closely connected to text complexity and by association to the age of the intended reader. For Dutch texts, 0.02 letters are added to words and sentences become 0.41 words longer when the readership ages with one year. In English, words lengthen with 0.006 letters and only 0.19 words are added to sentences in the same time frame. The effect of the age of the intended readership is thus smaller for the English part of the corpus than for the Dutch.

According to the statistical analyses conducted in this paper, there is a moderate correlation between titles of the same author when considering lexical diversity. Previous research indicates that lexical diversity is correlated with both the age of the author, where complexity decreases with age, and the age of the intended reader, where complexity increases with age. However, in

the first case, statistically there seems to be no significant effect between lexical diversity and the age of the author. The hypothesis does prove true, however, for the second case; there is a significant effect of the age of the intended reader as well as of language and the interaction between both features on the lexical diversity of texts in the corpus. When the LMM is split according to language, the model estimates that per year the intended reader of Dutch texts ages, the lexical diversity increases with 0.005941 units. This effect is smaller in the English texts included in the corpus, where the lexical diversity increases with only 0.004358 units. While the difference between these measurements seems small (0.001583), it is statistically significant for an average lexical diversity of 0.3037 for Dutch and 0.2883 for English texts. Thus, lexical diversity of Dutch texts is influenced by the age of the intended reader to a higher degree than English texts.

Statistically there is no significant effect of language, age of the intended reader or age of the author on the ratio between dialogue and narration. Furthermore, there is only a low correlation between measurements of titles of the same author.

4.7 Conclusion

While the textual features of average sentence length, average word length and lexical diversity are used both in authorship attribution and readability formulas, suggesting that writing style and text complexity are closely connected, a statistical analysis conducted on 692 Dutch and English texts written for children, young adults, and adults suggests that these elements are more related to categories based on the age of the intended reader. None of the features result in a strong correlation between titles of the same author. This supports the fact that the average sentence and word length have often been disputed as reliable for authorship attribution because they can be consciously manipulated by the author to produce a text with a desired level of readability in connection with the age of the intended reader of said text. However, lexical diversity, which is presented by some scholars as an element related to chronological issues, such as the age of the author, also turns out to be linked more closely to the categories determined by the age of the intended reader. The ratio between dialogue and narration, which was included in the analyses as a control feature, shows no correlation with any of the categories studied in this paper. In naive analyses using linear regression models, language has a statistically significant effect on all the textual features included in this study. However, the LMM which takes into account the author as a random effect refutes this conclusion; only the average sentence length is significantly influenced by the difference between Dutch and English texts. Haverals et al. took an important first step in gaining a deeper understanding of the interaction between the writing style and intended reader of a text. This paper built on that by showing that, next to function words, average sentence length, word length and lexical diversity are dependable features for readership attribution. Furthermore, it presented a method that is reliable for further singling out the set of features relevant to determine the age of the intended reader of a text.

[i] The analyses in this paper were conducted using the *lme4* package for R. Code and data repository found at: https://zenodo.org/record/7260676.

Chapter 5 Over (de) Grenzen

Op zoek naar de lezer in het oeuvre van Joke van Leeuwen[4]

Abstract – In the field of children's literature studies much attention has been devoted to analysing differences between children's and adult literature. Works by crosswriters, authors who write for both readerships in different works, are an excellent source for this research. This article adds to the debate by building upon previous studies which have used digital tools to track differences between literature for children and for adults. Using the oeuvre of Dutch crosswriter Joke van Leeuwen as a case study, it analyses formal and ideological aspects of each book. Although Van Leeuwen states that she does not see a difference in writing for either readership, this study finds formal differences in word and sentence length, lexical diversity and the ratio between dialogue and narrative. At the ideological level pertaining to perceptions of age, two recurrent themes are identified; imaging of the elderly and power relations between children and adults.

Context – Chapter five is based on an article published in a Dutch-language journal and revisits the analyses conducted in chapter three. Both focus on the oeuvre of just one author and consider formal textual elements. This chapter deconstructs the readability formulas included in the previous chapter and examines their most commonly used elements. Additionally, the age of the intended reader for the twenty-two titles examined in this chapter is much less contested. Although the size of the corpus might be considered small for traditional digital analyses, this chapter thus builds upon the suggestion for future research raised in chapter three. Together with the following chapter, the present acts as a bridge from one set of research questions to another; the analyses consider both the age constructed *by* the text (RQ 1.1), examined through formal features (RQ 2.2), and *in* the text (RQ 1.2), studied through the distribution of characters of different ages as well as explicit statements about age (RQ 2.4). For this wider array of analyses, the texts were used in two different versions as input; as raw text files and as annotated files. The examination of the set of formal features here led to the setup of chapter four, where the same features are examined on a larger scale. The analyses in this chapter are closest to close reading methods of all the publications included in this dissertation. This is partly because it reports on early research when more advanced digital methods were still being developed. On the other hand, the more limited corpus of this chapter opens up the opportunity to take some of the analyses back to the full text as well as the author's own views on writing for readers of different ages (RQ 2.6). The discussion of the metareflections often does this in order to study the context of these statements, an approach that is not used in any of the other chapters.

[4] This chapter is based on: Geybels, Lindsey. 'Over (de) grenzen: op zoek naar de lezer in het oeuvre van Joke van Leeuwen'. *Spiegel der Letteren*, vol. 63, no. 1–2, 2021, pp. 113–137.

5.1 Introductie

In geen enkel veld van literatuurstudie is er zoveel discussie gevoerd over het begrip 'grenzen', zowel in de betekenis van 'beperking' als van 'verdeling', als in het veld van jeugdliteratuur. Sommige literatuurwetenschappers stellen normen op voor de gepaste stijl, complexiteit en inhoud van kinderboeken. Over de status en toekomst van de grenzen tussen jeugd- en volwassenenliteratuur is er nog veel debat. Volgens Sandra Beckett (*Transcending Boundaries*) wordt deze onenigheid veroorzaakt door de opvallende veranderingen die de jeugdliteratuur in de laatste helft van de twintigste eeuw doormaakte en voornamelijk door de trend van de 'crossover'-literatuur in de jaren tachtig en negentig van de vorige eeuw. Deze werken, waarbij literatuur overgaat van een jeugdig naar een volwassen publiek of omgekeerd, dagen de traditionele definities van kinder- en jeugdliteratuur uit.

De beweging van het ene naar het andere publiek wordt door verschillende literatuurwetenschappers verbonden aan het concept van grenzen. Peter van den Hoven ('Grensverkeer') introduceerde de term 'grensverkeer' om te verwijzen naar de experimenten van verschillende Nederlandstalige auteurs met het schrijven voor lezers van een andere leeftijd aan het einde van de jaren tachtig. De bijdrage van Helma van Lierop-Debrauwer aan de uitgave *Transcending Boundaries* (Beckett, 1999), waarin zij de receptie van dubbelpublieksauteurs onderzoekt, is getiteld 'Crossing the Border'. Het omschrijven van de activiteit van zulke auteurs als het actief oversteken van de grens tussen jeugd- en volwassenenliteratuur, ongeacht de richting, impliceert dat er een grens aanwezig is. De aard en doorlaatbaarheid van die grens staan echter ter discussie in het zogenaamde grensverkeerdebat.

Dit artikel wil een bijdrage leveren aan dat debat via een onderzoek naar het oeuvre van Joke van Leeuwen, een Vlaams-Nederlandse auteur die schrijft voor verschillende leeftijden. Door het bestuderen van vormelijke en ideologische aspecten ga ik op zoek naar verschillen tussen haar werken op basis van de leeftijd van de beoogde lezer. Het onderzoek werd uitgevoerd met hulpmiddelen uit de digitale tekstanalyse. Naast de hoofdanalyse, onderzoek ik dus ook welke bijdrage kwantitatief onderzoek kan leveren aan een beter begrip van de verschillen tussen jeugd- en volwassenenliteratuur. Hiermee volgt dit artikel in de voetsporen van Helma Van Lierop-Debrauwer ('Grote Gelijkenis') en Nicole Hurkmans, die gebruik maakten van tekstanalyses gebaseerd op digitale principes voor het bestuderen van boeken geschreven voor verschillende leeftijden. Beiden focusten hun onderzoek op een vergelijking van stijl; zowel materiële aspecten, zoals lettergrootte of aantal woorden per pagina, als complexiteit, zoals zins- en woordlengte en enkelvoudige of samengestelde zinnen, werden onderzocht. Beiden onderzochten ook de verhouding tussen dialoog en beschrijving.

Tal van andere studies naar de verschillen tussen jeugd- en volwassenenliteratuur richtten zich op narratieve elementen, zoals stijl, genre en inhoud (zie o.a. Anderson, Beckett *Crossover Fiction*, Ghesquière en Talley). John Stephens stelt dat ook karakterisering beïnvloed wordt door de leeftijd van de beoogde lezer van een verhaal, "in het bijzonder de manier waarop opvattingen over kindertijd worden geconstrueerd door middel van karakterisering" (Stephens 'Maintaining', 184).[i] Dit aspect komt overeen met het gerepresenteerde kindbeeld, één van de elementen die Hurkmans noodzakelijk vindt in verder onderzoek naar de verschillen tussen jeugd- en volwassenenliteratuur.

Met hulpmiddelen uit de digitale tekstanalyse herneem ik in het eerste deel van dit artikel enkele vormelijke aspecten bestudeerd door Hurkmans en Van Lierop-Debrauwer, waaronder zinslengte en de verhouding tussen dialoog en beschrijvende tekst. Waar zij beiden echter slechts twee teksten tegenover elkaar zetten, analyseer ik 22 werken van Van Leeuwen. De vormelijke analyse wordt uitgebreid met een berekening van de lexicale diversiteit van elk boek, die een verdere aanduiding geeft van de vormelijke complexiteit. Het tweede deel is gewijd aan ideologische aspecten van de representatie van verschillende leeftijden. Zowel de grotere hoeveelheid teksten als het opnemen van volwassenenliteratuur stelt me in staat om niet alleen het kindbeeld te bestuderen, maar dat te relateren aan opvattingen over andere leeftijden. De ideologische opvattingen die kinderboeken impliciet, expliciet en via taal uitdrukken (Hollindale, *Ideology*), dragen bij tot hun opvoedende waarde. Het is daarbij de vraag of de opvattingen over leeftijd in de kinderboeken van Joke van Leeuwen verschillen van haar literatuur voor volwassenen. In eerste instantie zoek ik naar impliciete opvattingen over leeftijd, die tot uiting komen in de taal van personages van verschillende leeftijden. Een analyse van het aantal personages in elke leeftijdscategorie wordt vergeleken met het aantal woorden dat die personages uitspreken. Ten tweede onderzoek ik de opvattingen over leeftijd die expliciet zijn opgenomen in de teksten, in zogenaamde 'metareflecties' (Joosen *Adulthood*, 7).

5.2 Schrijven voor en over verschillende leeftijden

In 'Waar liggen de grenzen nog?' vergelijkt Hurkmans twee werken van Joke van Leeuwen: *Slopie* (2004), een jeugdboek, en *Vrije vormen* (2002), een boek voor volwassenen. Hurkmans gebruikt de term 'dubbelpublieksauteur' om te verwijzen naar Van Leeuwen en de drie andere auteurs die ze bestudeert. Deze benaming impliceert een tegenstelling tussen literatuur voor kinderen en volwassen, terwijl beide leeftijdscategorieën en ook de literatuur voor verschillende leeftijden zich eerder in een continuüm bevinden. De jeugdliteratuur is immers op haar beurt nog verder gesegmenteerd in verschillende leeftijdscategorieën, met adolescentenliteratuur als overgang tussen kinderboeken en boeken voor volwassenen. In dit artikel worden 22 werken van Van Leeuwen geanalyseerd en wordt er rekening gehouden met de diversiteit die bestaat binnen de begrensde categorieën van jeugd- en volwassenenliteratuur. De Engelstalige term 'crosswriter' heeft een bredere betekenis dan het Nederlandstalige 'dubbelpublieksauteur' en is dus beter geschikt voor auteurs zoals Van Leeuwen, die schrijven voor meer dan twee verschillende leeftijdscategorieën. Hoewel nog steeds omslachtiger dan het Engelstalige 'crosswriter', stel ik voor de Nederlandstalige term aan te passen naar 'multipublieksauteur' om te verwijzen naar auteurs die aparte werken adresseren aan lezers van meer dan twee verschillende leeftijden.

De term 'crosswriter' werd bedacht door U.C. Knoepflmacher in het begin van de jaren negentig en verwijst naar auteurs die kinderen en volwassenen in hetzelfde of in afzonderlijke werken aanspreken.[ii] Tot de jaren 1990 gaven auteurs er vaak niet veel ruchtbaarheid aan wanneer ze overstapten van volwassenen- naar jeugdliteratuur, en dat had vooral te maken met de relatief lage status van auteurs van jeugdliteratuur. Sommigen ontkenden hun functie als jeugdschrijver (Shavit *Poetics*, 39) of beweerden dat er op basis van leeftijd geen onderscheid bestaat in literatuur (Beckett *Transcending Boundaries*, xvii). Succesvolle auteurs als J.K. Rowling en Philip Pullman in het Verenigd Koninkrijk, en Bart Moeyaert en Joke van Leeuwen in Vlaanderen, hielpen om dit beeld bij te stellen.

Het oeuvre van Joke van Leeuwen bevat iets meer dan twintig prozateksten, opgedeeld als volgt: vijf boeken voor zevenjarige lezers, tien boeken voor negenjarigen, één boek voor twaalfjarigen en zes boeken voor volwassenen (zie Tabel 5.1).[iii] Dit artikel volgt de bepaling van de leeftijd van het doelpubliek zoals opgenomen in het Centraal Bestand Kinderboeken (CBK), een databank met bibliothecaire informatie van de meeste recente Nederlandstalige jeugdboeken. Voor alle boeken van Van Leeuwen vermeldt het CBK een leeftijdscategorie. In de analyses in dit artikel gebruik ik telkens de ondergrens, bijvoorbeeld 9 voor de categorie 9 tot 12.

Titel	Leeftijd geïntendeerde lezer
Alles nieuw (2008)	18
Bezoekjaren (1988)	12
De Appelmoesstraat is anders (1978)	7
Deesje (1985)	9
De metro van Magnus (1981)	9
De onervarenen (2015)	18
De tjilpmachine (1990)	18
Dit boek heet anders (1992)	9
Een huis met zeven kamers (1979)	7
Feest van het begin (2012)	18
Het verhaal van Bobbel (1987)	9
Hier (2018)	18
Iep! (1996)	9
Ik heet Reinier en ons huis is afgebrand (2020)	9
Kukel (1998)	9
Kweenie (2003)	7
Maar ik ben Frederik, zei Frederik (2013)	7
Slopie (2004)	9
Toen ik (2017)	7
Toen mijn vader een struik werd (2010)	9
Vrije vormen (2002)	18
Wijd weg (1991)	9

Tabel 5.1: Overzicht van de titels met de respectievelijke leeftijd van de beoogde lezer in het oeuvre van Joke van Leeuwen.

Van Leeuwen vertelde tijdens de boekvoorstelling van *Ik heet Reinier en ons huis is afgebrand* dat ze zelf niet graag een leeftijd op haar boeken plakt, omdat ze gelooft dat elk kind anders leest. Terwijl ze achter het idee staat van een minimumleeftijd voor boeken, zou er volgens haar op geen enkel boek een maximumleeftijd mogen staan. De Britse multipublieksauteur Philip Pullman heeft een vergelijkbare poëtica; hij gelooft dat schrijven voor kinderen en schrijven voor volwassenen alleen verschilt doordat de auteur niet kan verwachten dat een kind evenveel kennis heeft (Beckett *Crossover Fiction*, 118). Hij stelt dus een cognitieve ondergrens op literatuur, maar spreekt niet over een bovengrens. In lijn met de meeste multipublieksauteurs, pleit Joke van Leeuwen niet enkel langs de kant van de lezer voor het vervagen van de leeftijds-grenzen, ook voor zichzelf zegt ze geen verschil te zien in het schrijven voor kinderen of voor

volwassenen (Hurkmans, 108). De analyses op vormelijk en ideologisch vlak in dit artikel trachten aan te duiden dat er, in het geval van deze schrijver, op bepaalde vlakken wel verschillen zijn.

5.3 Op de grens tussen kwalitatief en kwantitatief onderzoek

Het meeste onderzoek naar de verschillen tussen jeugd- en volwassenenliteratuur is uitgevoerd door een gedetailleerde lezing van een klein corpus. Dit artikel stapt af van deze traditionele methode van literatuuronderzoek en doet beroep op technieken uit de digitale tekstanalyse (Digital Humanities). Hierdoor wordt het makkelijker om een groter aantal teksten tegelijk te bestuderen. Zowel Van Lierop-Debrauwer ('Grote Gelijkenis') als Hurkmans maakten al eerder gebruik van tekstanalyse om verschillen op te sporen tussen een jeugd- en een volwassenenboek, geschreven door dezelfde auteur. Beiden concluderen dat de teksten die geadresseerd zijn aan kinderen kortere zinnen en woorden bevatten en een groter percentage aan dialoog tegenover verhalende tekst in vergelijking met de boeken voor volwassenen. Hoewel beide analyses interessante inzichten opleveren, kan worden gesteld dat ze de gebruikte methoden niet ten volle hebben benut. Er is slechts een kleine selectie van pagina's onderzocht waarop enkel basis-methoden uit de tekstanalyse werden toegepast; Van Lierop-Debrauwer telt handmatig en Hurkmans gebruikt *Wordsmith*, een laagdrempelig computerprogramma voor tekstanalyse. Digitale tekstanalyse is beter geschikt voor het bestuderen van grotere tekstcorpora en zal dus in dit artikel gebruikt worden om de resultaten van deze steekproef te valideren, corrigeren en complementeren.

In *Introducing Electronic Text Analysis* somt Svenja Adolphs de belangrijkste voordelen op van het gebruik van digitale hulpmiddelen voor het bestuderen van taal en literatuur. Ten eerste resulteert het bestuderen van een groter aantal teksten in stabielere waarnemingen, omdat de invloed van afwijkingen, zoals atypische boektitels of personages, een kleiner gewicht in de resultaten zal hebben. Een tweede voordeel is een objectievere verwerking van de teksten. Bij het gebruik van computationele methoden om informatie uit teksten te halen, wordt de mogelijke invloed van de onderzoeker beperkt. Bovendien maakt het gebruik van instrumenten voor digitale tekstanalyse het onderzoeksproces reproduceerbaar, wat betekent dat het kan worden herhaald en geverifieerd door andere onderzoekers (7–8). Een belangrijke kanttekening is dat gesofisticeerde tekstanalyses nog steeds menselijke tussenkomst nodig hebben, bijvoorbeeld bij het instellen van parameterwaarden of het annoteren van tekstsegmenten. In dit artikel geef ik aan bij welke analyses dit het geval is, welke gevolgen verschillende instellingen kunnen hebben en welke waarden ik gebruik.

5.3.1 Vorm

Het bestuderen van de vormelijke aspecten van een tekst geeft inzicht in de moeilijkheidsgraad van die tekst. In het onderwijs wordt al decennia lang gebruik gemaakt van zogenaamde 'readability formulas', of formules voor leesbaarheid, om leesmateriaal voor kinderen te selecteren (Fry, 286). De numerieke score die deze formules opleveren, komt overeen met de leesvaardigheid van kinderen, die op haar beurt vaak gekoppeld wordt aan een specifieke leeftijdscategorie. De bekendste score in het Nederlands taalgebied is de AVI-score. Leesbaarheidsformules kunnen worden gebruikt om een minimale leeftijdsgrens te bepalen. In dit artikel zal ik echter geen gebruik maken van deze formules om twee redenen. Ten eerste bestaat er een handvol formules,[iv] waarvan de meeste licht afwijkende resultaten geven, en er is geen consen-

sus over welke het meest wenselijk is. Een vergelijking van de formules valt buiten het bestek van dit artikel, maar het feit dat er verschillen bestaan, doet af aan de autoriteit waarmee ze vaak gehanteerd worden. Ten tweede bevatten leesbaarheidsformules zo goed als allemaal dezelfde elementen (My Byline Media). De berekeningen bestaan uit een combinatie van woordlengte (in lettergrepen of karakters) en zinslengte, en in sommige gevallen het percentage van moeilijke woorden (bepaald door een vastgelegde lijst of door een minimum van drie lettergrepen). Elke formule maakt ook gebruik van numerieke variabelen, voor het bekomen van een genormaliseerd resultaat dat overeenstemt met het Amerikaanse onderwijssysteem. Aangezien het doel van dit artikel niet is de leeftijd van de beoogde lezer te bepalen op basis van de tekst, maar om eigenschappen te onderzoeken die verband houden met een reeds gedefinieerde leeftijd, analyseer ik in dit deel van het artikel de bouwstenen van deze formules in plaats van de formules zelf.[v]

Om de gemiddelde woordlengte te berekenen, wordt elke tekst ingelezen en gesplitst in afzonderlijke woorden waarbij alle leestekens en cijfers worden verwijderd. Van elk woord wordt het aantal karakters geteld, waarna het gemiddelde wordt berekend. Wanneer de teksten worden ingedeeld op basis van de leeftijd van de beoogde lezer (7, 9, 12 of 18), merken we twee dingen op: enerzijds is er slechts een klein verschil tussen de kortste (4,16) en langste (4,4) gemiddelde woordlengte, maar anderzijds is er wel een stijging naarmate de leeftijd van de beoogde lezer oploopt (zie Tabel 5.2). Van Leeuwen gebruikt soms erg lange woorden, maar opvallend genoeg komen die in zo goed als alle boeken voor. Het langste woord staat in *Alles nieuw* (2008), een boek voor volwassenen: Wettelijkeaansprakelijkheidsverzekeringsexpert (46 letters). Maar ook in *Slopie* (2004), een boek voor negenjarigen, vind je woorden als honderdvijfennegentigeneenhalfpoten (35 letters) en lezers van zeven jaar maken al kennis met betbetbetoverovergrootgrootvader (36 letters) in een boek als *Een huis met zeven kamers* (1979). In haar classificatie van neologismen in fictie, wijst Céline Poix de functie van 'attention-seeking devices' (ASDs), oftewel aandachtstrekkers, toe aan buitenproportioneel lange woorden. ASDs vestigen kortstondig de aandacht op het medium van de tekst in plaats van het verhaal dat de tekst communiceert. De neologismen hierboven zijn morfologisch van aard: ze zijn gecreëerd door de vorm van woorden te veranderen, in dit geval het samennemen van aparte woorden. Hoewel Poix neologismen bespreekt als kenmerkend voor jeugdliteratuur, keert deze vorm van poëtische vrijheid terug in het volledige oeuvre van Van Leeuwen.

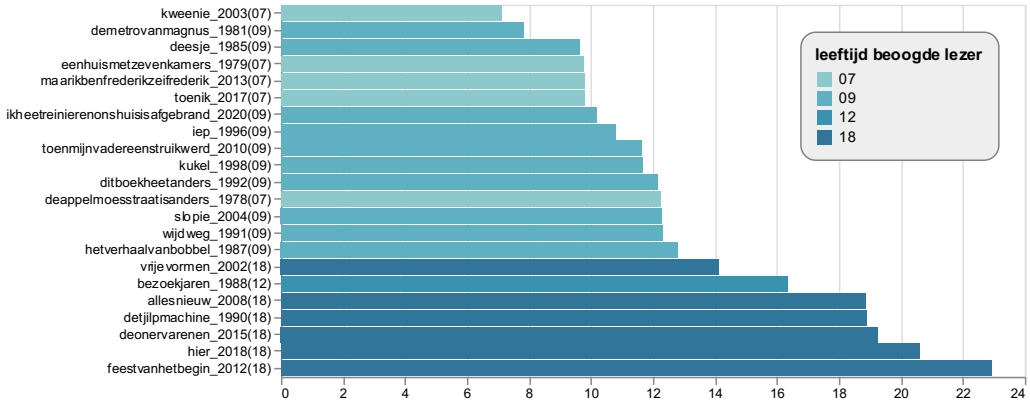
Minimumleeftijd beoogde lezer	7	9	12	18
Gemiddelde woordlengte (aantal karakters)	4,16	4,23	4,29	4,4

Tabel 5.2: Gemiddeld aantal letters per woord, ingedeeld per leeftijd van de beoogde lezer.

De waarden van de gemiddelde woordlengte per boek correleren sterk ($r = 0.77$)[vi] met de gemiddelde zinslengte; hoe langer de woorden in een tekst, hoe langer ook de zinnen. Volgens Colleen Lennon en Hal Burdick is "de beste voorspeller van de moeilijkheidsgraad van een zin de lengte ervan" (5).[vii] In een studie naar uitgeverijen van kinderboeken brengt Celia Catlett Anderson de lengte van zinnen in verband met de leeftijd van het lezerspubliek. Ze vroeg verschillende uitgeverijen naar het soort wijzigingen die ze voorstellen bij manuscripten. Eén van de meest voorkomende verzoeken is het inkorten van zinnen om zo de tekst simpeler te maken en toegankelijker voor jonge lezers (Anderson, 10). De gemiddelde zinslengte wordt in dit artikel

berekend door het totale aantal woorden te delen door het aantal zinnen, bepaald door een Python-functie uit de Natural Language Toolkit.

Figuur 5.1 toont de gemiddelde zinslengte van de 22 onderzochte romans in dit artikel in oplopende volgorde, variërend van 7,1 tot 22,9 woorden per zin (af te lezen op de x-as). De leeftijd van de beoogde lezer per boek is terug te vinden in de titel op de y-as en af te leiden uit de kleur van de staven: hoe donkerder de kleur, hoe ouder het lezerspubliek. De grafiek toont dat de groepering van de meeste teksten parallel loopt met de verschillende leeftijdscategorieën. Met uitzondering van *De metro van Magnus* (1981, 9+), *Deesje* (1985, 9+), *De Appelmoesstraat is anders* (1978, 7+) en *Vrije vormen* (2002, 18+) neemt de gemiddelde zinslengte toe met de leeftijd. Opmerkelijk zijn de hoge scores voor de volwassenenromans. Dit is te wijten aan de invloed van enkele erg lange zinnen. In *Feest van het begin* (2012, 18+), de roman met het hoogste gemiddelde aantal woorden per zin, vind je zinnen met maar liefst 245, 203 en 193 woorden. Deze drie zinnen, eveneens als de langste zinnen uit de andere volwassenenromans, bevatten elk een combinatie van indirecte rede en vrije indirecte rede van één of meerdere personages. Hoewel deze buitensporige waarden het gemiddelde erg naar boven trekken, krijgen we een gelijkaardig beeld bij het berekenen van de mediaan van de zinslengtes van elk boek. Het enige verschil is dat de mediaan van *De Appelmoesstraat is anders* (1978, 7+) zich minder uitzonderlijk hoog verhoudt tegenover de waarden van andere boeken in dezelfde leeftijdscategorie.

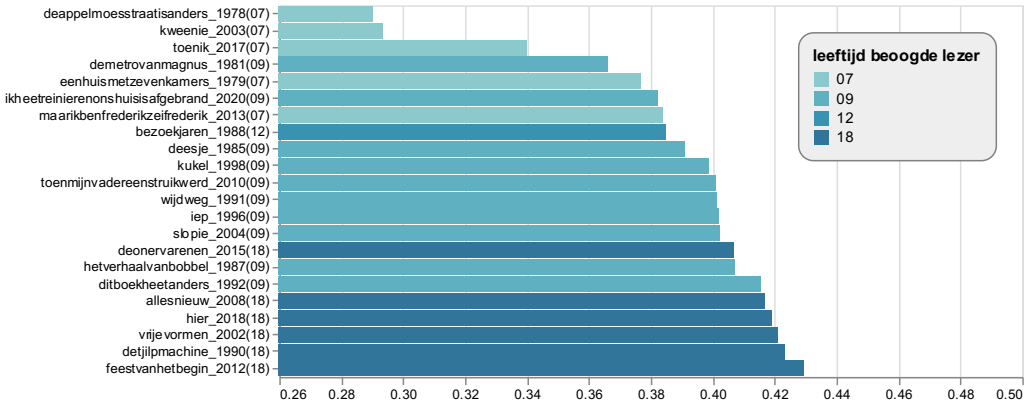


Figuur 5.1: Gemiddeld aantal woorden per zin, ingedeeld per leeftijd van de beoogde lezer.

Hoewel de bovenstaande analyse de observaties van Van Lierop-Debrauwer ('Grote Gelijkenis') en Hurkmans bevestigen, is het nuttig om dieper in te gaan op de stilistische verschillen tussen literatuur voor verschillende leeftijden. Op basis van de vaststelling van Victoria Johansson dat "lexicale diversiteit een goede maatstaf is om verschillen tussen leeftijdsgroepen op te sporen" (77),[viii] gebruik ik het om de moeilijkheidsgraad van Van Leeuwens boeken verder te bestuderen. Lexicale diversiteit is een voorstelling van het aantal verschillende woorden in een tekst, meestal weergegeven door de 'type-token'-verhouding (TTR), de verhouding tussen het aantal unieke woorden ('types') en het totale aantal woorden ('tokens') (Johnson). Een welbekend probleem van deze verhouding is dat het resultaat erg onderhevig is aan de lengte van de tekst; hoe langer een tekst, hoe kleiner de kans dat er een nieuw uniek woord zal verschijnen. Om dit vervormend effect van de tekstlengte te voorkomen, hebben Michael Covington en Joe McFall de Moving Average Type-Token Ratio (MATTR) ontwikkeld, een methode voor het berekenen van de type-token-verhouding waarbij een venster van een bepaald aantal woorden wordt genomen, de TTR van deze selectie berekend wordt en vervolgens het venster een woord naar voren schuift om

de TTR opnieuw te berekenen en zo verder. Na het doorlopen van de gehele tekst wordt het gemiddelde van alle TTR-scores berekend. De grootte van het venster is een van de parameters die door de onderzoeker worden bepaald en die de resultaten van de analyse kan beïnvloeden.

Covington en McFall adviseren een venstergrootte van 500 woorden bij het uitvoeren van een stylometrische analyse. In de analyses van dit artikel heb ik de grootte van het venster ingesteld op 1000. Deze waarde ligt dichter bij het totale aantal woorden van het kortste boek in het corpus, *De Appelmoesstraat is anders* (1978, 1419 woorden). Het doel van de huidige analyse is immers het opsporen van de omvang van de woordenschat die nodig is om de tekst te begrijpen. We verwachten dat teksten bedoeld voor jongere lezers meer herhalingen bevatten en dus een beperktere woordenschat hebben (Bland, 156). Figuur 5.2 toont aan dat de meeste boeken in het corpus deze hypothese bevestigen. De algemene trend is: hoe jonger het beoogde lezerspubliek, hoe lager de lexicale diversiteit van het boek. Er zijn echter enkele uitzonderingen; *De metro van Magnus* (1981, 9+) wordt opnieuw eerder bij de categorie van boeken voor zevenjarigen gegroepeerd dan bij de titels voor negenjarigen. Terwijl de lexicale diversiteit van *Maar ik ben Frederik, zei Frederik* (2013, 7+) hoger ligt dan het gemiddelde voor zevenjarigen, bevindt *Ik heet Reinier en ons huis is afgebrand* (2020, 9+), het meest recente jeugdboek van Van Leeuwen, zich aan de lage kant voor een boek gericht aan negenjarigen. *Bezoekjaren* (1988), als enige boek voor twaalfjarigen heeft een opvallend lage MATTR-score, en past op vlak van lexicale diversiteit eerder tussen de boeken voor zeven- en negenjarigen.[ix] Ook in de overgang tussen boeken voor negenjarigen en volwassenen zien we een zone waarin de tendens doorbroken wordt; de score van *De onervarenen* (2015, 18+) strookt meer met de lexicale diversiteit van boeken voor negenjarigen, terwijl *Dit boek heet anders* (1992, 9+) aanleunt tegen boeken voor volwassenen. De analyse van lexicale diversiteit geeft een gevarieerder beeld van de verschillen tussen boeken voor verschillende leeftijden dan de analyses van woord- en zinslengte. De grenzen tussen de leeftijdscategorieën lijken bij lexicale diversiteit minder strikt: we zien eerder bredere overgangsgebieden.

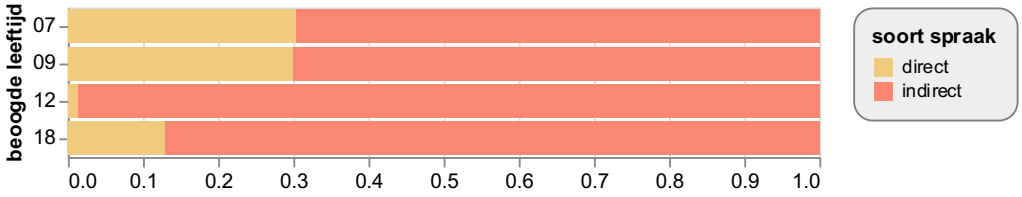


Figuur 5.2: MATTR-score berekend met een venstergrootte van 1000 woorden, ingedeeld per leeftijd van de beoogde lezer.

Volgens Rita Ghesquière wordt ook de verhouding tussen dialoog en beschrijvingen beïnvloed door de leeftijd van de beoogde lezer: boeken voor jongere lezers bevatten doorgaans meer dialoog (96). Deze observatie wordt bevestigd door Hurkmans' analyse van boeken geschreven door dubbelpublieksauteurs. Ze vergeleek Van Leeuwens volwassenenroman *Vrije vormen* (2002) met *Slopie* (2004), een boek gericht op negenjarigen, en vond 39% directe rede in het

kinderboek versus 23% in de volwassenenroman. Hoewel er digitale hulpmiddelen beschikbaar zijn die met een vrij hoge accuraatheid directe van indirecte rede kunnen onderscheiden op basis van aanhalingstekens, koos ik voor deze analyse voor een handmatige identificatie van spraak. Sommige teksten van Van Leeuwen bevatten immers geen aanhalingstekens[x] en het identificeren en toekennen van spraak aan een specifiek personage is nodig voor de analyse in het volgende deel van dit artikel.[xi] Met behulp van een zelfgeschreven Python-script wordt alle directe rede geïdentificeerd en kan die apart van indirecte rede worden geanalyseerd.

Het hier bestudeerde corpus toont een verschil dat vergelijkbaar is met de waarneming van Hurkmans, zoals blijkt uit Figuur 5.3. De boeken zijn gegroepeerd per leeftijdscategorie (7, 9, 12 en 18), weergegeven op de y-as van de grafiek. Op de horizontale as is de verhouding tussen directe en indirecte rede voorgesteld door het aantal woorden in elke soort spraak procentueel te visualiseren. Het aantal woorden in de directe rede, dus het aandeel van de dialogen, daalt wanneer de leeftijd van de beoogde lezer stijgt; in boeken voor zeven- en negenjarigen beslaat het gemiddeld 30% van het totale aantal woorden, terwijl dit slechts 13% is in boeken gericht op volwassenen. Wanneer we naar de boeken afzonderlijk kijken, wijken enkele titels af van het gemiddelde van hun leeftijdscategorie: *Toen ik* (2017, 7+, 16%), *Ik heet Reinier en ons huis is afgebrand* (2020, 9+, 8%), *De metro van Magnus* (1981, 9+, 40%), *De tjilpmachine* (1990, 18+, 35%) en *De onervarenen* (2015, 18+, 3%). De verhouding tussen dialoog en narratief in de categorie van twaalfjarigen wijkt af van de gemiddelde observatie. Gezien er slechts één boek is opgenomen in deze categorie, wordt dit resultaat sterk beïnvloed door de stilistische keuze van de auteur om in dit verhaal de spraak van personages zo goed als nooit in directe rede te noteren.



Figuur 5.3: De verhouding tussen dialoog en narratief, ingedeeld per leeftijd van de beoogde lezer.

Uit de digitale analyse van vormelijke aspecten in het werk van Joke van Leeuwen blijkt een duidelijk verschil tussen de jeugd- en volwassenenboeken. Met enkele uitzonderingen is er zelfs een evolutie waarneembaar in de jeugdboeken naarmate de leeftijd van de beoogde lezer stijgt. Aanvullend aan de ideologie die vervat zit in de taalvaardigheid bestudeerd in de vormelijke analyse, worden in het volgende deel van dit artikel impliciete en expliciete uitingen van ideologie bestudeerd, meer bepaald het beeld dat gevormd wordt van verschillende leeftijden. Hierbij kijken we niet langer naar vormelijke verschillen in de boeken die samenhangen met de leeftijd van het doelpubliek, maar naar de constructie van leeftijd op het niveau van de verteller en de personages.

5.3.2 Impliciete ideologie

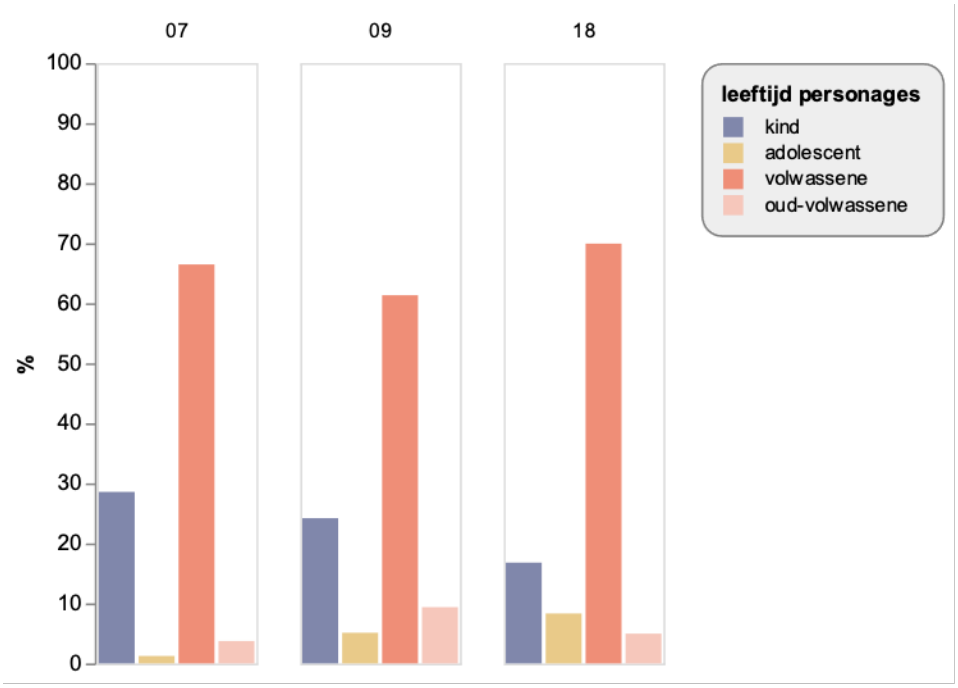
De verdeling van de directe rede over verschillende leeftijdscategorieën kan bijdragen tot de ideologische opvattingen over leeftijd. Kinderboeken bieden een grote verscheidenheid aan opvattingen over leeftijd aan jonge lezers, wat bijdraagt tot de pedagogische en socialiserende waarde die traditioneel toegekend wordt aan deze literatuur. Om te onderzoeken of deze functie

afneemt naarmate de leeftijd van de beoogde lezer stijgt, bestuderen we de ideologische aspecten waarbij we inzoomen op de boeken voor zeven- en negenjarigen en volwassenen. Het boek voor 12+ laten we hier buiten beschouwing omdat het slechts om één titel gaat. Eerst proberen we te bepalen welke leeftijdscategorieën een stem krijgen in het verhaal en hoe groot dat aandeel is. In haar studie naar de narratieve constructie van personages, drukt Maria Nikolajeva op het belang om te bepalen wie spreekt en waarschuwt ze voor de mogelijke invloed van de vertelinstantie op de manier waarop de lezer de spraak van personages interpreteert (*Rhetoric*, 239–40). Digitale tekstanalyse biedt op dit vlak opnieuw een voordeel; door indirecte rede, dat waardeoordelen over uitspraken van personages bevat, buiten beschouwing te laten, wordt de analyse niet beïnvloed door een, volgens Nikolajeva veelal volwassen, vertelinstantie.

Voor de analyse verzamelen we eerst informatie over alle personages die aanwezig zijn in het verhaal. Tijdens het manueel annoteren worden alle personages geïdentificeerd die minimaal twee keer vernoemd worden. Zo worden figuren zoals een onbelangrijke passant buiten beschouwing gelaten in de analyses. Alle personages worden opgedeeld in vier leeftijdscategorieën, die zijn afgeleid van onderzoek in de leeftijdsstudies zoals Lorraine Green en Thomas Armstrong: 'kind' (0–11 jaar), 'adolescent' (12–19 jaar), 'volwassene' (20–59 jaar) en 'oudere' (60 jaar en ouder). De personages worden zo nauwkeurig mogelijk toegewezen aan één van deze categorieën op basis van expliciete vermeldingen van leeftijd of van informatie uit de context. Een personage dat op de lagere school zit, wordt bijvoorbeeld als kind opgenomen; een personage dat werkt, wordt als volwassene geïnterpreteerd. In verschillende boeken van Joke van Leeuwen vinden we personages die verouderen: in *Bezoekjaren* (1988) en *Feest van het begin* (2012) bijvoorbeeld volgen we de hoofdpersonages gedurende verschillende fases in hun leven, en samen met hen worden ook de andere personages ouder. We deelden deze verouderende personages op in verschillende categorieën al naargelang hun leeftijd op dat moment. Personages waarover de tekst geen informatie bevat op vlak van leeftijd, bijvoorbeeld een anonieme groep mensen op straat, worden toegewezen aan een vijfde categorie: 'leeftijdloos'. De noodzaak van een manuele annotatie voor het traceren van de specifieke leeftijd van een personage draagt bij tot het argument dat digitale tekstanalyse vaak niet mogelijk is zonder een diepere lezing van de teksten in kwestie.

Figuur 5.4 bevat drie staafdiagrammen, één voor elke groep teksten gebaseerd op de leeftijd van de beoogde lezer. Elke grafiek visualiseert het percentage aan personages onderverdeeld volgens de vier leeftijdscategorieën. Opvallend is het hoge aandeel van volwassen personages in alle teksten, wat mogelijk indruist tegen de verwachting dat jeugdliteratuur vooral over jonge personages gaat; in elke groep teksten behoren minstens 60% van de geïdentificeerde personages tot de leeftijdscategorie 'volwassene'. Toch mogen we hier niet automatisch uit afleiden dat de boeken de volwassenheid van deze personages ook thematiseren. Wanneer we kijken naar de functies die deze personages vervullen binnen het verhaal,[xii] dan is het duidelijk dat het merendeel van deze groep zogenaamde 'backdrop characters', of achtergrondfiguren, zijn (Nikolajeva *Rhetoric*, 113–14): personages die worden toegevoegd om het verhaal realistischer te maken en voornamelijk gekarakteriseerd worden door het beroep dat ze uitoefenen, bijvoorbeeld 'receptioniste' of 'instrumentmaker'. Bijgevolg zijn deze soort personages meestal volwassenen. Bij de categorieën van kind- en adolescente personages zien we een tendens die omgekeerd evenredig is aan de stijgende leeftijd. Het aandeel kindpersonages neemt af van 28% in boeken voor 7-jarigen naar 24% in boeken voor 9-jarigen en verder naar 17% in volwassenenboeken. Het

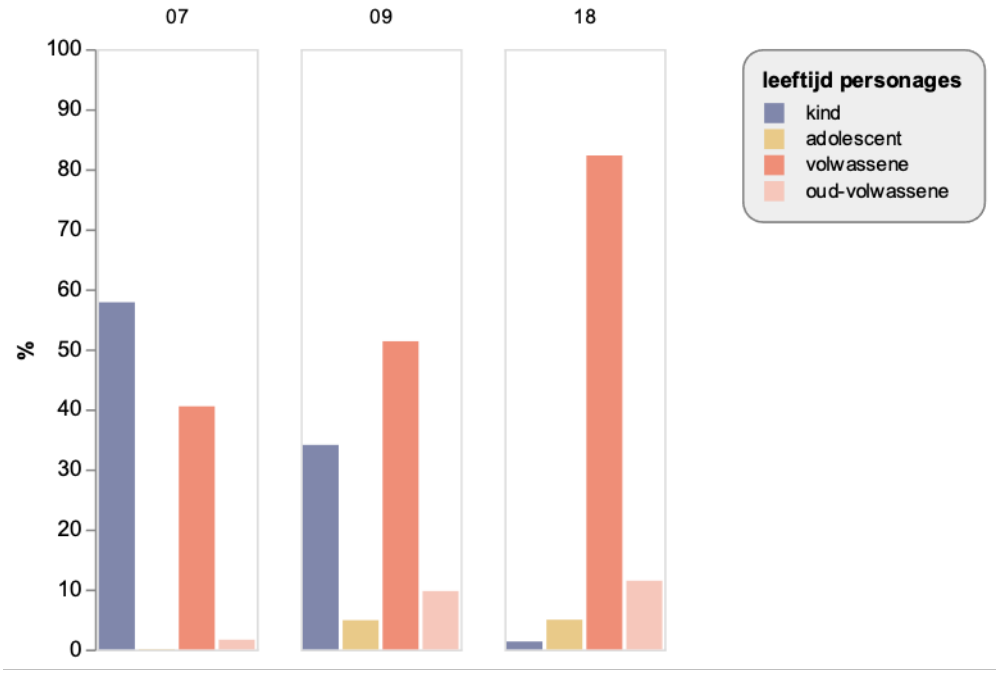
aandeel van adolescenten kent een kleinere variatie; deze waarde neemt toe van 1,3% naar 5% en voorts naar 8%.



Figuur 5.4: Percentage van personages per leeftijdscategorie, onderverdeeld volgens leeftijd van de beoogde lezer.

Op zich zijn de bovenstaande statistieken niet zo veelzeggend omdat ze beïnvloed worden door personages die soms een kleine rol spelen in het verhaal. De directe rede geeft meer informatie over het aandeel van personages van verschillende leeftijdscategorieën in het verhaal. Figuur 5.5 geeft een beeld van het relatieve aantal woorden in de directe rede van personages ingedeeld volgens de vier leeftijdscategorieën, opnieuw voor elke groep teksten. Hoewel in boeken voor zevenjarigen voornamelijk volwassen personages voorkomen, worden de meeste woorden uitgesproken door kinderen. Adolescenten en oud-volwassenen krijgen amper het woord in deze boeken. Dit wijst op een focus op de relatie tussen kinderen en volwassenen, waar kinderen, tenminste op vlak van de representatie van stem, de bovenhand krijgen. Een bijkomstige verklaring voor de hoge aanwezigheid van volwassenen, maar een kleiner aandeel in de directe rede, is het grote aantal aan 'backdrop characters'. Zoals eerder vermeld, valt het op dat die vaak volwassen zijn.

Opvallend is dat volwassenen in de boeken voor negenjarigen de overhand krijgen in de directe rede. Dit kan deels te wijten zijn aan het kleinere aantal kinderen in combinatie met prominentere rollen voor volwassen personages. Bovendien bevestigt dit beeld de indruk van Vanessa Joosen (*Adulthood*) dat jeugdliteratuur niet alleen kindbeelden aan de lezers biedt, maar ook een rijke bron is voor fictieve constructies van volwassenheid. Omgekeerd is de rol van kinderen in de boeken voor volwassenen beperkt, althans als je afgaat op hun aandeel in de directe rede. In de boeken voor volwassenen zijn gemiddeld 17% van de personages kinderen, maar hun spraak beslaat amper 1,3%. Zowel de aanwezigheid van oud-volwassenen als hun aandeel in de directe rede is over de hele lijn klein, en in de boeken voor de jongste lezers komen er zelden oudere personages aan het woord. Hiermee onderscheidt het werk van Joke van Leeuwen zich van andere jeugdboeken, waar grootouders bijvoorbeeld vaak wel een prominente rol spelen.



Figuur 5.5: Percentage van woorden in spraak van personages per leeftijdscategorie, onderverdeeld volgens leeftijd van de beoogde lezer.

5.3.3 Expliciete ideeën over leeftijd

De voorgaande analyses over de aanwezigheid en spraak van personages van verschillende leeftijdscategorieën zijn gebaseerd op impliciete opvattingen over de ideologie van leeftijd. Om een meer betrouwbaar beeld te presenteren over hoe deze ideologie vervat zit in het werk van Van Leeuwen, meer specifiek welke verschillen er zijn tussen boeken bedoeld voor lezers van verschillende leeftijden, bespreek ik in dit deel van het artikel expliciete uitspraken over leeftijd in haar werk. Deze metareflecties zijn voorbeelden van leeftijdsnormen, meestal impliciet vervat in de maatschappij maar ook expliciet vastgelegd in de wet, die in de tekst geëxpliciteerd worden. Deze leeftijdsnormen zijn sociale constructies en dus zeer afhankelijk van de cultuur waarin ze functioneren. Er bestaan bijvoorbeeld expliciete normen over de leeftijd waarop je een auto mag besturen en impliciete normen over wanneer je te oud bent om te huilen.[xiii]

De analyse van die geëxpliceerde leeftijdsnormen is de minst digitale van aard in dit artikel. Het verzamelen van de metareflecties gebeurde op dezelfde wijze als het identificeren van spraak: we annoteren expliciete uitspraken over leeftijd en kunnen die informatie koppelen aan de personages (en hun leeftijd) die de uitspraken doen.[xiv] Via een zelfgeschreven Python-script kunnen we zo snel een overzicht creëren van de leeftijdsnormen die personages per leeftijds-categorie expliciteren. Het analyseren van deze uitspraken gebeurt in dit artikel niet digitaal, maar aan de hand van traditionelere onderzoekstechnieken, namelijk het identificeren en bespreken van opvallende thema's. De reden hiervoor is de beperkte data die de metareflecties uit het oeuvre van Van Leeuwen aanleveren (68 metareflecties uit 17 boeken). Wanneer een grotere dataset beschikbaar is, zou de toepassing van sentimentanalyse zeer zinvol zijn. Zo zouden inzichten bekomen kunnen worden over hoe volwassenen zich positioneren tegenover kinderen of wat adolescenten denken over ouderen. Bovendien is het belangrijk om de meta-reflecties in de juiste context te lezen: soms zijn ze gekleurd door ironie, overdrijving of door het negatieve licht waarin het verhaal het sprekende personage plaatst.

Na het verzamelen van alle metareflecties die geïdentificeerd werden, valt meteen op dat het aantal metareflecties op leeftijd verhoudingsgewijs afneemt naarmate de leeftijd van de beoogde lezer stijgt. In de boeken voor zevenjarigen komen gemiddeld drie metareflecties voor op tienduizend woorden. Voor negenjarigen zijn dit er nog maar twee en de boeken voor volwassenen tellen amper één metareflectie per tienduizend woorden.[xv] Dit is een tendens die we ook bij andere auteurs zien en die de indruk bevestigt dat leeftijd en intergenerationele verhoudingen belangrijkere thema's zijn in de jeugdliteratuur dan in literatuur voor volwassenen. Om dit verder te onderzoeken, bekijken we de inhoud van de reflecties. De omvang van dit artikel maakt een uitgebreide analyse van elke metareflectie niet mogelijk. In plaats daarvan ligt de focus op eventuele verschillen die zich voordoen in de implicaties van de uitspraken tussen boeken voor verschillende leeftijden. Aangezien er slechts twee metareflecties zijn over adolescenten (één in een boek voor negenjarigen en één in een volwassenenroman), worden ze buiten beschouwing gelaten.[xvi] Voorts zijn er 31 metareflecties over kinderen, 6 over volwassenen, 22 over oude mensen en 7 uitspraken waarin meer dan één leeftijdscategorie voorkomt. Verhoudingsgewijs worden de meeste uitspraken over kindertijd gemaakt in boeken voor zevenjarigen (7/11). Inhoudelijk zijn er enkele thema's te onderscheiden in de metareflecties, waarvan de twee meest prominente hier besproken zullen worden: de machtsrelatie tussen kinderen en volwassenen en de beeldvorming van oudere mensen.

Zowel in boeken bedoeld voor zeven- en negenjarigen als in Van Leeuwens volwassenenromans is de verhouding tussen kinderen en volwassenen een terugkerend thema. *Maar ik ben Frederik, zei Frederik* (2013, 7+) thematiseert wat het betekent om volwassen dan wel kind te zijn. Het is een verhaal over een volwassen man die midden in de dag een plotse leeftijdsverandering doormaakt: hij zit vast in het lichaam van een jongetje. Hoewel deze metamorfose minder gruwelijk is dan die van Gregor Samsa in de roman van Franz Kafka, bevat *Frederik* eveneens een aantal reflecties over de verandering en de nieuwe gedaante van het hoofdpersonage. Doordat het hoofdpersonage drastisch verjongt, terwijl hij zijn herinneringen behoudt, gaan zo goed als alle metareflecties uit dit boek over het feit dat kinderen *te jong* zijn om een heleboel dingen te doen. Handelingen die voor Frederik als volwassen man vanzelfsprekend waren, worden hem nu ontzegd. Zo krijgt hij bijvoorbeeld te horen, steevast van volwassenen personages, dat hij zeker nog geen kranten leest, niet alleen hoort te leven, laat staan een eigen huis kan hebben, en veel te jong is om achter het stuur te zitten of een rijbewijs te hebben. Frederik beseft na een tijd dat zijn transformatie hem op sommige vlakken machteloos maakt: "Kon hij hem maar tegenhouden. Maar dat kon hij niet. Hij was te klein".[xvii]

Het beoordelen van kinderen op basis van de competenties van volwassenen, en specifiek het beschouwen van volwassenheid als de norm, wordt door Nikolajeva aangeduid als 'aetonormativity', analoog aan de term heteronormatief in genderstudies (2010). Het volledige verhaal van *Maar ik ben Frederik, zei Frederik* (2013) geeft echter een meer genuanceerd beeld op de machtsverhouding die wordt gesuggereerd door de metareflecties. Van Leeuwen benoemt ook regelmatig gelijkenissen tussen kinderen en volwassenen en het verhaal komt tot een goed einde enkel door een samenwerking en gelijke verstandhouding tussen jong en oud. Dit is, volgens Van Leeuwen zelf, één van de troeven van haar verhalen en iets waar ze telkens naar streeft: een verbinding tussen verschillende leeftijden, niet enkel in de personages die ze creëert, maar ook in de lezers van haar boeken ('Ik heet Reinier').

In boeken voor negenjarigen wordt deze opzet al iets meer vertegenwoordigd in de metareflecties, hoewel het merendeel opnieuw reflecteert over het (on)vermogen geassocieerd met een jonge leeftijd. Kinderen zijn volgens volwassenen personages of de vertelinstantie "niet sterk genoeg om te weten dat je niet alles en iedereen moet geloven" (*Dit boek heet anders*, 1992), "weten nog niet wat liegen en verzwijgen is" (*Ik heet Reinier en ons huis is afgebrand*, 2020) en zijn niet in staat om een kasteel te bestormen (*Kukel*, 1998). Twee boeken bevatten echter elk een metareflectie die in contrast staat tegenover het thema van het kind als machteloze persoon en benadrukken dat kinderen over bepaalde rechten beschikken omwille van hun jonge leeftijd. Soms is dat ludiek – een volwassen personage stelt dat "[e]lk kind recht heeft op een schoolreisje" (*Dit boek heet anders*, 1992) – soms serieuzer, bijvoorbeeld wanneer de vader van Reinier met een metafoor duidelijk maakt dat je als kind recht hebt om omringd te zijn met liefde (*Ik heet Reinier en ons huis is afgebrand*, 2020).

In *Dit boek heet anders* (1992, 9+) volgt een handvol metareflecties elkaar kort op in een scène waarin gelijkenissen tussen kinderen en volwassenen lijnrecht tegenover de machtsrelatie tussen beiden wordt gezet. In deze scène discussieert het hoofdpersonage, het kind Wammie, met een volwassen personage, meneer Hijdaar, over of volwassenen, al dan niet kinderen, te veel lawaai maken. Wanneer Wammie vraagt waarom er geen kinderen aanwezig zijn in het deftige restaurant waar het gesprek plaatsvindt, antwoordt meneer Hijdaar dat "ze vinden dat die soms wat veel lawaai maken". Wammie schiet meteen in de verdediging en zegt dat volwassenen veel meer lawaai maken "en die stinken veel harder". Meneer Hijdaar probeert Wammie te sussen en zegt diplomatisch: "kinderen en volwassenen maken volgens mij evenveel lawaai. Maar niet altijd op dezelfde plaatsen en tijden". Hij erkent de karaktereigenschap van beide leeftijdsgroepen zonder negatief te spreken over één van beide. Voor Wammie daarentegen is het hek van de dam: "[volwassenen] roepen tegen kinderen dat ze zachter moeten doen, en zelf blijven ze lawaai maken. Kinderen moeten altijd doen wat grote mensen zeggen". Wammie's zusje, Diddie, besluit het gesprek met te zeggen dat "grote mensen vaak niet zo goed luisteren [naar kinderen]". Hoewel meneer Hijdaar moeite doet om Wammie's negatieve perceptie van volwassenen bij te stellen, bevestigt hij de machtspositie van volwassenen over kinderen door het gesprek te beëindigen wanneer hij zijn gelijk niet kan halen.

In boeken voor volwassenen komt de nadruk meer te liggen op de gelijkenissen van kinderen en volwassenen en vormen de metareflecties een positiever kindbeeld. Zo wordt de boodschap meegegeven dat volwassenen kinderen vaker serieus moeten nemen in de volgende vaststelling: "Een kind begrijpt vaak meer dan het zeggen kan" (*De onervarenen*, 2015). In de roman *Vrije vormen* (2002) wordt er tweemaal naar kinderen verwezen om het gedrag van volwassenen te beschrijven en zo beide leeftijdscategorieën te verbinden: een vriend van het hoofdpersonage "omhelst haar met al zijn armen, zoals een kind zijn moeder" en eerder in het verhaal kruipt een man "als een jongen over de vloer" bij het opmeten van een kamer. Bovendien wordt er in deze roman gereflecteerd op de speelsheid van kinderen, als eigenschap die volwassenen niet langer bezitten. Terwijl dit opnieuw een verschil tussen kinderen en volwassenen aanhaalt, weerspreekt dit idee de overwegende opvatting uit de kinderboeken dat kinderen verschillende eigenschappen van volwassenen nog niet bezitten. In een gedachtenexperiment beelden twee twintigers zich in hoe kinderen en volwassenen zouden reageren op een zitbank die in het midden van een wandelpad in het park geplaatst wordt:

> [Z]e denken dat volwassenen zullen aarzelen om erop te gaan zitten, zo midden op het wandelpad, maar kinderen zullen erover willen lopen van de ene kant naar de andere, en als het een dag is waarop veel vaders komen die hun kind een weekend mogen houden voor het weer terug moet naar de moeder, dan zullen die vaders geduldig hun kind vasthouden en hen wel vier keer heen en weer laten lopen op die bank voor ze zeggen dat het genoeg is en dat heen en weer lopen niet opschiet. (*Vrije vormen*, 2002)

Een tweede thema dat de metareflecties in de boeken van Van Leeuwen naar voren brengt, is de beeldvorming van oudere mensen. Waar in het vorige thema kinderen vaak lijnrecht tegenover volwassenen werden geplaatst, worden in een aantal boeken voor negenjarigen kindertijd en ouderdom met elkaar verbonden, een wederkerend motief in jeugdliteratuur. Dit motief vervalt vaak in de stereotype opvatting van de 'infantilised senior', de oud-volwassene die wordt afgeschilderd als kind. In deze opvatting worden de "beperkingen van de kindertijd vergeleken met de gebreken van ouderdom" (Hockey en James, 137).[xviii] In *Wijd weg* (1991) wordt dit stereotype beeld op impliciete wijze opgeroepen wanneer gesteld wordt dat "oude handen" geen lepeltjes meer kunnen vasthouden, iets wat ook voor jonge kinderen geldt. In hetzelfde boek worden beide leeftijdscategorieën ook explicieter verbonden wanneer de levensloop van de mens wordt voorgesteld als volgt: "Je begon als kind in een klein kamertje, midden in je volwassen leven had je kans op iets groters, en daarna moest je het maar weer met minder doen" (*Wijd weg*, 1991).

In Van Leeuwens jeugdboeken komt naast de verbinding tussen kinderen en ouderen, ook de dood aan bod, als een normaal aspect van een ouder leven. De jonge protagonist en verteller in *Ik heet Reinier en ons huis is afgebrand* (2020) stelt dat "opa's nu eenmaal doodgaan". Een ander kindpersonage merkt op dat het normaal is om geen ouders meer te hebben wanneer je oud bent (*Slopie*, 2004). Niet enkel kinderen reflecteren over de sterfelijkheid van oude volwassenen; in *De metro van Magnus* (1981) geeft de grootmoeder van Magnus aan dat ze bijna klaar is met leven en oud genoeg is om dood te gaan.

In de volwassenenromans wordt er vaker gereflecteerd op oude volwassenen dan een andere leeftijd. Dit geeft aan dat leeftijdscategorieën met een klein aandeel in de directe rede toch een belangrijke component van het verhaal kunnen vormen. De meeste reflecties komen uit *Alles nieuw* (2008), een verhaal waarin het hoofdpersonage veel tijd spendeert bij een vrouw van in de tachtig. Vooral de fysieke en mentale kenmerken van de oudere leeftijd worden benoemd, zoals snel bang zijn, rimpels en een hangende buik hebben, het niet meer kunnen inschatten van de leeftijd van mensen een stuk jonger dan jezelf en het begin van dementie. Parallel daaraan zijn een aantal metareflecties die, net zoals in de jeugdboeken, het beeld van de geïnfantiliseerde oudere oproepen. De jongvolwassen protagonist van het verhaal vraagt zich bijvoorbeeld af "hoe het zou zijn als iemand [je lichaam] professioneel waste omdat je dat zelf niet meer kon". Verder door brengt ze ouderen rechtstreeks in verband met kinderen:

> Ze waren allemaal op de laatste bladzijden van lange levens, maar lazen die nauwelijks van elkaar. Ze aten en maakten geluid. Ze waren elkaars gezelschap omdat ze hun ouderdom deelden, zoals kinderen op een lagere school op basis van hun leeftijd bij elkaar zaten, niet omdat ze dezelfde hobby hadden of dezelfde sport beoefenden. (*Alles nieuw*, 2008)

De boeken van Van Leeuwen die gericht zijn op lezers van zeven en negen jaar drukken een welbekende frustratie uit van kinderen die vaak te horen krijgen dat ze te jong zijn om iets te doen, zien of begrijpen. In de boeken voor negenjarigen, waar tevens enkele voordelen van kindertijd

worden belicht, wordt dit licht genuanceerd. In volwassenromans verschuift de machtspositie van volwassenen over kinderen nog verder. Hier ligt de nadruk meer op de gelijkenissen tussen kinderen en volwassenen en de troeven van kindertijd, zoals speelsheid, die volwassenen verliezen. Het beeld dat gevormd wordt van oudere volwassenen loopt gelijk tussen Van Leeuwens jeugdboeken en boeken voor volwassenen: dat van de geïnfantiliseerde oudere. Waar de jeugdboeken daarbovenop mortaliteit ter sprake brengen, richten de volwassenromans zich eerder tot de fysieke en mentale kenmerken van ouderdom.

5.4 Conclusie

Bij de resultaten van dit artikel moet opgemerkt worden dat het doelpubliek dat vermeld wordt in het CBK niet noodzakelijk bepaald is door de auteur, of alleen door de auteur. De uitgeverij heeft hierin meestal ook een hand. Bovendien hoeven de stilistische en ideologische aspecten niet noodzakelijk toegeschreven te worden aan een bewuste auteursintentie. Dat gezegd zijnde, heeft het gebruik van digitale instrumenten om het oeuvre van multipublieksauteur Joke van Leeuwen te analyseren een aantal interessante observaties aan het licht gebracht. Hoewel ze beweert dat zij geen verschil ziet tussen schrijven voor volwassenen en schrijven voor kinderen, tonen de voorgaande analyses aan dat er wel degelijk op sommige vlakken een onderscheid gemaakt kan worden tussen haar boeken voor verschillende beoogde lezers. Voornamelijk op het vlak van stijl doet zich een duidelijk onderscheid voor. Op enkele uitzonderingen na stijgen zowel de woord- en zinslengte als de lexicale diversiteit samen met de leeftijd van de beoogde lezer. De verhouding tussen indirecte rede en dialoog maakt een tweescheiding tussen haar werk voor kinderen en volwassenen.

De ideologische aspecten die impliciet worden uitgedrukt in de teksten, werden bestudeerd door de aanwezigheid van personages uit de vier verschillende leeftijdscategorieën en het aandeel van de spraak van deze personages in het totale aantal woorden in de directe rede. Naarmate de leeftijd van het doelpubliek stijgt, is er een duidelijke afname van het aandeel kindpersonages. Nog markanter is de afname van het aandeel kinderenspraak in de directe rede, terwijl volwassenpersonages steeds meer praten naarmate de leeftijd van het doelpubliek stijgt. Terwijl kinderen de luidste stem krijgen in jeugdboeken, worden ze tot zwijgen gebracht in de boeken voor volwassenen.

In de analyse van de expliciete uitingen van de ideologie rond leeftijd werden enkele beperkingen die digitale analyses met zich meebrengen duidelijk. Ten eerste kunnen de volledige troeven van een digitale analyse niet worden uitgespeeld wanneer er niet voldoende data is. Dit werd zowel uit een bespreking van de studies van Hurkmans en Van Lierop-Debrauwer duidelijk als uit de analyse van metareflecties in het laatste deel van dit artikel. Ten tweede zien we enkele discrepanties opduiken tussen de digitale analyse en het onderzoek naar de leeftijdsnormen die geëxpliciteerd worden in de teksten. Hoewel oudere personages minder aanwezig zijn, zowel in het aantal personages als in hun aandeel in de directe rede, worden er een groot aantal metareflecties op ouderdom opgenomen in de teksten.

Een derde manier waarop de expliciete uitingen van ideologie het beeld van de kwantitatieve analyse bijstellen is het kleinere bewijs van verschil tussen de verschillende leeftijden van de beoogde lezer; waar woord- en zinslengte en lexicale diversiteit de teksten opdeelden volgens dit criterium, blijft voornamelijk het beeld dat gevormd wordt van ouderen quasi gelijk voor alle

boeken. Wel valt de machtsverschuiving tussen kinderen en volwassenen op. Als laatste is het belangrijk op te merken dat, door de metareflecties met digitale hulpmiddelen te onttrekken uit de tekst, er onvermijdelijk een deel van de context verloren gaat, bijvoorbeeld de rol die het personage dat de uitspraak doet vervult. Vaak wordt er gebruik gemaakt van een negatief personage om een reflectie op een omstreden uitspraak uit te lokken bij de lezer en het de bedoeling is dat deze tot een tegengestelde conclusie komt. Uit deze discrepanties blijkt het belang van het complementeren van digitale analyses met traditionele narratieve analyses.

[i] Deze en de volgende vertalingen zijn van mijn hand. Oorspronkelijke tekst: “specifically in the way in which notions of childhood are constructed through characterisation” (Stephens ‘Maintaining’, 184).

[ii] Beckett brengt haar gebruik van de term terug naar een sessie die U.C. Knoepflmacher leidde tijdens de Modern Language Convention van 1993 (Beckett *Transcending Boundaries*, xi-xii).

[iii] Boeken voor eerste lezers werden niet opgenomen in de primaire literatuur van dit artikel, omdat deze lectuur stilistisch zeer specifiek is en vaak onderhevig is aan regels opgelegd door de uitgeverij. Voorts werden verzamelingen van kortverhalen buiten beschouwing gelaten.

[iv] De formules die het meest gebruikt worden in studies naar de leeftijd van de beoogde lezer zijn o.a. Automated Readability Index (ARI), Gunning Fog, Dale-Chall, Simple Measure of Gobbledygook (SMOG) en Flesh-Kincaid.

[v] Alle boeken die in dit artikel geanalyseerd worden, zijn verkregen via uitgeverij Querido en de DBNL met toestemming van de auteur, of werden eigenhandig gescand en gedigitaliseerd. De analyses zijn uitgevoerd met behulp van zelfgeschreven scripts in de programmeertaal Python.

[vi] Waarde van de Pearsons correlatiecoëfficiënt, een maat, tussen 0 en 1, die de sterkte van de relatie tussen twee continue variabelen weergeeft.

[vii] Oorspronkelijke tekst: “the best predictor of the difficulty of a sentence is its length” (Lennon and Burdick, 5).

[viii] Oorspronkelijke tekst: “lexical diversity is a better measure to use for detecting differences between age groups” (Johansson, 7).

[ix] Hoewel dit boek atypisch is op vlak van de leeftijd van de beoogde lezer, maakt het integraal deel uit van Van Leeuwens oeuvre en werd het daarom opgenomen in de analyses.

[x] *De onervarenen* (2015) gebruikt bijvoorbeeld geen aanduiding en in sommige gevallen een liggend streepje aan het begin van een nieuwe regel.

[xi] De teksten werden geannoteerd aan de hand van richtlijnen opgesteld door het Tekst Encoding Initiative (TEI).

[xii] Deze informatie is mee opgenomen in de externe lijst met metadata over de personages.

[xiii] Beide voorbeelden zijn afgeleid uit metareflecties gevonden in boeken van Van Leeuwen:
1) “Je bent veel te jong om achter het stuur te zitten en je hebt natuurlijk ook geen rijbewijs” (*Maar ik ben Frederik, zei Frederik*, 2013).
2) “Ik had grote mensen haast nooit zien huilen” (*Toen ik*, 2017).

[xiv] Aangezien het opsporen van deze uitspraken meer onderhevig is aan interpretatie van de annotator dan de andere twee elementen, benadruk ik nogmaals dat alle werken die geanalyseerd worden in dit artikel geannoteerd werden door dezelfde persoon.

[xv] Deze getallen zijn een relatieve voorstelling van het aantal metareflecties per groep boeken per leeftijdscategorie. Het totaal aantal woorden per groep werd omgerekend naar tienduizend voor een duidelijke vergelijking tussen de leeftijdscategorieën.

[xvi] “Hij moest wel een jaar of vier ouder zijn dan zij. Hij had al een groefje tussen zijn ogen” (*Wijd weg*, 1991). Het vrouwelijke personage hier is twaalf jaar oud. De metareflectie impliceert dus dat adolescenten jonger dan 16 jaar geen groefje tussen de ogen hebben.
“[...] maar toen ze weer weg wilde lopen, hoorde ze plotseling een stem uit de kast, een stem die zich nog niet had genesteld, als van een jongen tussen kind en man” (*De tjilpmachine*, 1990).

[xvii] De analyses in dit artikel werden uitgevoerd op digitale teksten. Hierdoor worden er geen paginanummers vermeld bij citaten.

[xviii] Oorspronkelijke tekst: “the apparent ‘limitations’ of childhood are mapped on to a parallel series of ‘inadequacies’ believed to characterise old age” (Hockey en James, 137).

Chapter 6 "Weird, but lovely"

A Digital Exploration of Age in David Almond's Oeuvre[5]

Abstract – This chapter relies on tools from digital humanities to gain a deeper understanding of fictional age and the age of the intended reader in the oeuvre of David Almond by studying a large part of this oeuvre simultaneously. Starting off as an author for adults, he never intended to write for a younger readership before publishing *Skellig*. This raises the question whether there are differences between his books intended for readers of different ages. In the first part of this chapter, Almond's writing style is studied using stylometric analyses to uncover clustering according to the age of the intended reader. The second part uses two different types of analyses to explore ideologies in terms of age contained within Almond's fictional characters. First, topic modelling is applied to the speech of characters belonging to specific age groups to study the themes that these characters address. Second, a syntactic parser is used to extract the adjectives, verbs and possessions associated with different age groups. By combining the observations from the three analyses, this chapter highlights the assets digital tools have in terms of reproducibility, objectivity, and uncovering general trends in a large body of texts when employed in the study of age. The efficiency of digital tools when it comes to a large dataset also makes it possible to conduct a variety of analyses in a short span of time to construct a comprehensive picture of this aspect in the oeuvre of David Almond.

Context – Like chapters three and five, this chapter focuses on the oeuvre of a single author and, like the previous chapter, contributes to answering RQ 1.1 (how does the age of the intended reader affect the construction of age *by* a work of fiction?) and RQ 1.2 (how does the age of the intended reader affect the construction of age *in* a work of fiction?). The difference with the previous chapter is the selection of analyses; while the previous chapter used basic digital tools, this chapter employs various established methods within the field of digital humanities to construct an elaborate study of age. Due to the use of different types of analysis, each of which looks at the construction of age in a different way, the conclusion might not be as coherent as those of the chapters that have a narrower focus. To study differences between the texts as a whole, stylometric analyses are conducted on the unannotated texts (RQ 2.1). This method was applied to the works David Almond along with those of other British authors in chapter two, but there, only the striking writing style of his children's book *Klaus Vogel and the Bad Lads* was discussed. Age *in* the text is investigated on the level of fictional characters in two ways: first, by analysing the speech of characters of different ages and, second, using a syntactic parser to study words associated with characters (RQ 2.5). For these analyses, the annotated versions of the texts are used. This chapter is part of co-authored monograph on methods to analyse the construction of age in David Almond's oeuvre and includes references to the other chapters in that book which might seem out of place for a stand-alone publication. To tie in with findings

[5] This chapter is based on: Geybels, Lindsey. '"Weird, but lovely": A Digital Exploration of Age in David Almond's Oeuvre'. *Constructing Age for Young Readers: An Exploration of Methods for the Analysis of Age in David Almond's Work*, Routledge, 2023, pp. 59–91.

from other studies reported in the edited volume, some of the results in this chapter are brought back to Almond's views on writing for readers of different ages (RQ 2.6).

6.1 Introduction

Infants dream of monsters,
the young dream dreams of love,
the old dream dreams of being young.
(Almond *A Song for Ella Grey*, 201)

Such is the uncomplicated view of adolescent character Orpheus on differences between people of various ages in *A Song for Ella Grey* (2014), David Almond's young adult novel based on the myth of Orpheus and Eurydice. In truth, age, both in real life and in fiction, is a much more complex concept. Do infants truly have the imagination to dream up monsters, are old people actually that nostalgic, and where are the younger adults in this line? Do they not dream at all? These contrasting views on characters of different ages, as well as Almond's status as a cross-writer, raise the question to what extent his books published for readers of different ages vary, both in their writing style and in the age norms they convey.

In this chapter, I demonstrate the potential benefit of an intersection between the field of digital humanities and the study of children's literature by exploring various facets of the representation of age in Almond's work through three different digital tools: stylometry, topic modelling and syntactic parsing. The stylometric analysis lays bare possible distinctions in the writing style of texts intended for readers of different ages. Topic modelling and syntactic parsing are conducted to explore how ideology pertaining to age is represented in Almond's works in both explicit and implicit ways, through the content of character speech and characterisation as it is established through verbs, adjectives and possessions associated with characters of different ages. To get a comprehensive image of the different facets of the representation of age in Almond's oeuvre, the analyses in this chapter are performed on a large portion of his works, namely, 21 titles. By studying such a large number of texts at once, the following discussion supplements research on Almond that has emerged in recent decades by moving away from close-reading strategies and towards studying an extensive corpus by using tools from the field of digital humanities. While this kind of research has long been limited to adult fiction, various researchers in children's literature studies have recently started using a distant reading approach (van Lierop-Debrauwer 'Grote Gelijkenis', Hurkmans, Karsdorp, Cross et al., Haverals et al., Haverals and Joosen, Giddens, and Fitzsimmons and Alteri).

As established in the previous chapters, Almond is a 'polygraphic' author who publishes books for readers of different ages and so breaks through the dichotomy implied by traditional views on crosswriting as moving from an adult to a child readership or vice versa. Moreover, the analyses take into account that children's literature itself is further segmented into different age categories, from picturebooks to young adult literature. Almond's books have been published for children of various ages. The analyses in this chapter classify his work into three categories: children's literature for ages seven to eleven, young adult literature, ranging from twelve to seventeen, and adult literature. In reality, these age categories are situated on a continuum rather than being opposites, but since digital analyses rely on countable categories, I grouped them together according to the most common segmentation as it is also found in library catalogues and bookstores. The combined observations from the analyses paint a comprehensive

picture of the representation of age in the oeuvre of David Almond and reveal stylistic and ideological differences according to the age of the intended reader.

In *Introducing Electronic Text Analysis* (2006), Svenja Adolphs lists the main advantages of using digital tools to study language and literature (7–8). First, digital analyses allow researchers to study a large number of texts simultaneously, and reveal trends that operate across a large corpus; smaller-scale studies run the risk of giving more weight to atypical observations, that such digital analyses can avoid. A second advantage that Adolphs mentions, is a more objective processing of the texts by reducing the possible subjectivity of the researcher. However, it must be noted that in analyses like those included in this chapter, the influence of the researcher still exists in the annotation and interpretation of the resulting data. Third, the use of tools for digital text analysis makes the research process reproducible, meaning that it can be repeated and verified by other scholars. While I cannot share the annotated texts for reasons of copyright, I have included some of the results of my digital analyses in their entirety in the text of this chapter and in the appendices. That way, readers can add their own interpretation of the data that my digital analyses yielded. Before those analyses are presented, I will give a brief overview of the study of writing style and of characterisation as it is studied in children's literature and digital humanities. After that, I will explain the digital methodologies employed in this chapter.

6.2 Studying writing style through digital analyses

The differences and overlaps between children's and adult literature have received ample attention in children's literature studies but have rarely been studied using quantitative analyses. Comparisons of writing style have so far been mainly limited to investigations into the complexity of texts, with children's literature being defined in rather vague terms like 'clarity', 'directness', 'simplicity' and an overall 'childlike style' (see Anderson, Ghesquière, Wall, Beckett *Transcending Boundaries* and *Crossover Fiction*). Ruth Bottigheimer observed that there exists a link between the literary style of a text and the reader that text is addressed to: "If a text was intended for children rather than for adults […] then both content and style should betray the presence, and identity, of those readers" (197). The stylometric analyses included in this chapter test this claim, using Almond as a case study.

During the past few decades, computational methods to determine the complexity of texts have appeared; for example, readability formulas that are mainly applied to children's literature in a pedagogical context. Stylometry provides additional digital tools to study the writing style of a text, but it has yet to be applied to children's literature on a large scale.[i] Together with Wouter Haverals and Vanessa Joosen, I investigated the applicability of stylometric methods not for authorship attribution but for stylistic clustering according to the age of the author and the age of the intended reader with promising results for the latter. We found that cluster analyses based on a small set of most frequent words, which in effect consists of function words, reveal that texts often cross the boundaries of authorship to group together according to the age of the intended reader. This means that books published for the same age range are stylistically more similar than texts published by the same author. In this chapter, I will apply a similar method to the works of David Almond.

6.3 Studying characterisation through digital analyses

Traditionally, characterisation is studied mainly using close reading strategies, but in recent years, many studies have employed digital tools to look into various aspects of fictional characters, including their social networks (see Oelke et al., Rochat and Kaplan, Kim and Klinger, and Thomas 'Long Novels'), minds (Mahlberg et al.), emotions (Nalisnick and Baird, and Jacobs) and proper names (van Dalen-Oskam and van Zundert, and Rowe et al.). Fictional characters, which are such an important aspect of most narrative forms, have long been neglected by children's literature studies (Nikolajeva *Rhetoric*, vii) but have received much more attention since the turn of the century.[ii] Maria Nikolajeva's *The Rhetoric of Character in Children's Literature* (2002) has become a cornerstone to study characters in fiction for young readers. In this narratological study, Nikolajeva situates different ways of characterisation on a scale from direct to indirect methods, rather than contrasting these two categories. At the far end of the scale, description and narration are the most direct methods of characterisation. The scale then moves through action and events, speech, and internal representation to implicit characterisation, such as proper names and settings (157–58). To study the construction of age through characters in Almond's novels, two out of the three analyses in this chapter are designed to represent this scale and thus include methods from both ends as well as the middle. For explicit characterisation, I study description and action through adjectives and verbs attributed to characters, respectively. To study implicit characterisation, I look at the possessions of characters. Possessions are here understood in a grammatical sense: that is as attributes linked to characters through possessive forms (genitives, possessive pronouns). Although Nikolajeva only considers permanent attributes as a method of characterisation, the analysis below broadens this category to include all objects that belong to a character, from their clothes to their dreams. Another analysis looks at speech, which Nikolajeva places in the middle of the scale. Whereas she considers it in essence to be a more direct method, in practice she finds that in children's literature, speech is mainly used to drive the plot rather than to develop characters.[iii]

6.4 Data collection

The analyses that follow incorporate the majority of Almond's oeuvre, as the results of digital analyses become more meaningful the more extensive the studied material becomes. However, this entails a trade-off concerning the qualitative aspect of the analysis; it will lack the in-depth nature and nuances presented by traditional close reading analyses and focuses more on general trends and attitudes in how age is represented in Almond's work. Although working with digital tools creates the opportunity to study a large body of texts, this way of conducting research requires a critical approach to which types of text are included as primary data. To avoid that the results of the analyses are influenced by differences on a textual level, short story collections, picture books, primers, graphic novels, plays and the novel *The True Tale of the Monster Billy Dean* (2011) are not included in the corpus.[iv] This book is Almond's first full-length novel, published simultaneously for a young adult and adult audience and written entirely in an idiosyncratic language. The story is told in the voice of the protagonist Billy Dean, who has grown up in a highly secluded environment and has received minimal education. His story is written down in phonetic language, for example: "This tail is told by 1 that died at birth by 1 that came into the world in days of endles war and at the moment of disaster. He grew in isolayshon wile the enjins of destrucshon flew and smoke rose over the sitys and wile wilderness and waste crept all acros the world" (1). Due to the different spelling in this book compared with the rest of the corpus,

The True Tale would be anomalous in an analysis of writing style and characters' speech and contribute little when looking at the descriptions, actions and possessions of characters.

Table 6.1 provides an overview of the 21 titles that are included in the analyses in this chapter, along with the age of the intended reader for each work. While many texts are written, published or marketed with the age of a readership in mind, the age of the intended reader can be conceptualised in specific or broad terms. This information is not always communicated publicly, especially for books published in the UK. This is partly because of writers' opposition against 'book banding', a strategy of children's literature publishers to increase sales by adding age recommendations to the covers to assist adult buyers in their search for suitable reading material for their children. In 2008, many authors, led by best-selling Philip Pullman, argued that such banding would have the opposite effect of restricting readers in their choice of books (The Page Turner; see also Pullman 'No to Age Banding'). None of David Almond's works that I examine include such an age marker on the book itself. However, information about the age of the intended readership can still be found in the catalogues of certain publishers or on the websites of authors who are in favour of this method to classify children's literature. To categorise the books below, I first looked for information on the age of the intended reader on the websites of publishers and Almond's own website. When no minimum age or age range was given there, I consulted websites of booksellers and the Dutch database for children's books (CBK), which also

Title	Age of the intended reader
Harry Miller's Run (2015)	7
My Dad's a Birdman (2007)	8
The Boy Who Climbed into the Moon (2010)	8
The Boy Who Swam with Piranhas (2012)	8
Brand New Boy (2020)	8
The Tale of Angelino Brown (2017)	9
War is Over (2018)	9
Skellig (1998)	12
Kit's Wilderness (1999)	12
Counting Stars (2000)	12
Heaven Eyes (2000)	12
Secret Heart (2001)	12
The Fire-Eaters (2003)	12
Klaus Vogel and the Bad Lads (2009)	12
My Name is Mina (2010)	12
A Song For Ella Grey (2014)	12
The Colour of the Sun (2018)	12
Bone Music (2021)	12
Clay (2005)	15
Jackdaw Summer (2008)	15
The Tightrope Walkers (2014)	18

Table 6.1: List of novels by David Almond included in the corpus of this chapter and the corresponding age of the intended reader.

includes entries for many English children's books. For the analyses in the next sections of this chapter, the novels are placed into three broader categories: adult novels (eighteen and up), young adult fiction (twelve to fifteen) and children's literature (seven to nine).

To conduct digital analyses on the 21 texts, each work was digitised, either through obtaining an e-book version or scanning the physical book and running it through Optical Character Recognition (OCR) software. All texts were pre-processed to remove metadata, including the front and back matter, page numbers and chapter headings. For the second set of analyses in this chapter, those pertaining to the age of fictional characters, information is added during a process called 'text annotation', which means that words and (parts of) sentences of interest are identified and tagged. For this purpose, Extensive Mark-up Language (XML) is used in accordance with guidelines set up by the Text Encoding Initiative, which provides researchers with a "standard for the representation of texts in digital form" (TEI Consortium) using a hierarchical system of so-called tags. These tags can be used to categorise parts of a text, such as chapter headings or copyright information, or to add information, such as a character's name to a pronoun referring to that character.

The annotations add information about fictional characters in several ways. First, each text is divided into narration (indirect speech) and character speech (direct speech).[v] In case of direct speech, the tag includes an attribute to identify the speaking character. The following fragment shows an annotated version of the opening lines of *My Dad's a Birdman* (2007):

<said direct="false">An ordinary spring morning in 12 Lark Lane. The birds were tweeting and whistling outside. The city traffic rumbled and roared. Lizzie's alarm went ringa-ding-ding.</said> […]

<said direct="true" who="lizzie">"Dad!"</said>

<said direct="false">she shouted.</said>

<said direct="true" who="lizzie">"Daddy!"</said>

<said direct="false">No answer.</said>

In the case of third-person narration, <said direct="false"> is used to mark the beginning of parts of the text told by the narrator, while <said direct="true" who="lizzie"> marks the beginning of direct speech by a character, in this case, Lizzie. </said> is a so-called 'closing tag', marked by the slash after the first angle bracket. It is used to indicate where a section of direct or indirect speech ends. In the tag that marks direct speech, 'lizzie' is a so-called 'character ID'. The character IDs used in the annotated document are saved in a separate data frame, together with additional features, including the full name, species, gender and age of each character. We call this data frame the 'character list'. Below, I elaborate on this part of the annotation process.

In addition to direct and indirect speech, all references to characters are tagged: names, nouns and pronouns.[vi] The above fragment then looks like this:

<said direct="false">An ordinary spring morning in 12 Lark Lane. The birds were tweeting and whistling outside. The city traffic rumbled and roared. <rs ref="lizzie">Lizzie's</rs> alarm went ringa-ding-ding.</said> […]

<said direct="true" who="lizzie">"<rs ref="dad">Dad</rs>!"</said>

<said direct="false"><rs ref="lizzie">she</rs> shouted.</said>

<said direct="true" who="lizzie">"<rs ref="dad">Daddy</rs>!"</said>

<said direct="false">No answer.</said>

<rs ref="lizzie"> labels the words referring to characters with the right character ID. Again, this tag comes with a closing tag to mark the end of the reference: </rs>.

As explained above, the annotators make a separate data frame (e.g., an Excel table) that we call the 'character list', where they match the character ID with the character's name, species, ethnicity, gender and age (see Table 6.2 for a simplified example). The latter is recorded either in a numerical format or as a life stage, following a scheme (Table 6.3) based on research by life course scholars Lorraine Green and Thomas Armstrong. Age is a dynamic concept in a story; a character may age during narratives that span several years (e.g., J.K. Rowling's *Harry Potter* series), and many stories contain flashbacks in which the characters are younger than in the main part of the story. For this reason, a singular character may have several character IDs, each reflecting a different age. Similar to determining the age of the intended reader for Almond's work, identifying the age of his fictional characters is not without its difficulties or ambiguities. When assigning an age to human characters, the annotator either relies on explicit age markers or context; parents of minors are labelled adults in the broad sense, grandparents are categorised as old adults and those attending secondary school are considered to be adolescents. While this method of assigning age does not correctly reflect the true diversity of age, where people under twenty years of age can be parents or have a profession, performing digital analyses on a large amount of data sometimes requires such oversimplifications. The downside of this approach is that the age norms that are to be studied are in part also being used to create the annotated material, risking the creation of biased data.

Character ID	Full name	Gender	Relation to protagonist	Age	Life stage
carlo	Carlo Brooks	male	friend		adolescent
claire	Claire Wilkinson	female	friend		adolescent
claire17	Claire Wilkinson	female	friend	17	
claire5	Claire Wilkinson	female	friend	5	
ella	Ella Grey	female	protagonist		adolescent
fatherella	Grey	male	parent		adult

Table 6.2: Simplified example of a character list from the annotation of A Song For Ella Grey.

Life stage	Corresponding numerical age
infant	0–2
child	3–11
● earlychild	3–5
● middlechild	6–8
● latechild	9–11
adolescent	12–19
adult	20–…
● earlyadult	20–39
● middleadult	40–59
● oldadult	60–79
● deepoldadult	80–…

Table 6.3: Age scheme including the life stages and their corresponding numerical age used in the annotations of David Almond's oeuvre.

A further obstacle when assigning age to Almond's characters is the fantastical element in his work. Many characters are realistic portrayals resembling real people, but his stories also regularly feature ambiguous or supernatural creatures, a characteristic of the magic realist style which is often attributed to Almond's work (Latham). Michelle Anya Anjirbag and Frauke Pauwels note that the movie adaptation of *Skellig* hints at the age of its title character by portraying him as 'someone older'. However, in the novel the only clue to Skellig's age is that he is referred to as a 'man' on several occasions, suggesting that he is an adult. What complicates the attribution of age to Skellig is that he is revealed to be a supernatural creature that unites aspects of a human being and an angel. This raises the question whether life stages catered to human development are applicable to non-human creatures. A similar dilemma occurred for the supernatural creation in *Clay* that Emma-Louise Silva discusses in a chapter on social and material minds. For this reason, supernatural creatures and animated objects are not included in the analyses on the representation of age in the following section.

6.5 Style and the age of the intended reader

The analyses in this section use digital methods to investigate if David Almond's writing style differs in texts published for different ages. A small set of most frequent words, which are in effect function words, are used to perform the analyses in this chapter. The use of function words as discriminators for an author's writing style has been widely researched and confirmed (see Burrows 'Delta', Diederich et al., Grieve, Hoover 'Multivariate', Koppel et al. 'Author Gender', Martindale and McKenzie, Uzuner and Katz, and Yu). The raw digital text material of the 21 books in the corpus was subjected to a second round of pre-processing, after the removal of metadata. Below is a list of the 100 most frequent words extracted from the samples created from Almond's works that were used for the stylometric analysis:

about, across, again, all, an, and, #are, around, as, at, away, aye, back, #be, #been, but, by, #came, #can, #come, #could, #day, #did, #do, down, each, #eyes, #face, for, from, #get, #go, good, #got, #had,

#hand, #have, #head, here, how, if, in, into, #is, just, #know, #like, little, #ll, #look, #looked, #man, more, no, not, now, of, off, on, one, or, other, out, over, #put, #re, right, #said, #saw, #say, #says, #see, so, some, still, #tell, that, the, then, there, #things, #think, this, through, #time, to, up, #ve, #was, #way, #went, #were, when, where, #whispered, #will, with, #world, #would, yes

After the list of most frequent words is manually edited to remove, amongst others, verbs that might be influenced by the narrative point of view or tense, we can start measuring how closely related the different texts are.[vii] The stylistic analysis is presented in two forms: a hierarchical cluster analysis (HCA) and a principal component analysis (PCA).[viii] The HCA (Figure 6.1) clusters stylistically similar documents together in a tree-shaped diagram. The larger the distance between two documents, or clusters of documents, the smaller their similarity. The PCA (Figure 6.2) is a scatterplot, where each dot represents one sample of Almond's texts. Here too, stylistically similar samples are positioned together, and the larger the distance between two dots, the smaller their similarity. To highlight the feature of the age of the intended reader in the HCA and PCA graphs, the labels in both visualisations are colour-coded. The samples of texts aimed at children range from orange for seven-year olds, to yellow for readers aged eight years old, green for nine-year olds, to blue and purple for young adult titles (twelve and fifteen years old respectively), while red titles signify adult novels.

Almond believes that his writing changed between his short stories for adults and working on Skellig (Almond, Interview with David Almond), which suggests a potential stylometric difference between texts intended for readers of different ages. The HCA tree diagram in Figure 6.1 displays evidence of this, as various texts for similar ages cluster together. However, this grouping is not perfect, and his writing style does not seem to be solely influenced by the intended reader. Moreover, the two analyses show a slightly different pattern. In the HCA tree diagram (Figure 6.1), Skellig clusters closest together with The Fire-Eaters and Heaven Eyes, two other novels intended for twelve-year olds, but can also be considered to be quite similar to his adult novel The Tightrope Walkers, as they are connected in the hierarchical tree only one level higher (at distance 8). However, in the PCA (Figure 6.2) both titles are located on opposite sides of the vertical axis, suggesting stylometric distinction; the samples that make up Skellig are found at the top of the plot while the red dots that represent The Tightrope Walkers are at the bottom. The PCA scatterplot (Figure 6.2) shows a larger cluster of novels for twelve-year olds than the HCA tree diagram. Located in the upper left quadrant, represented by blue dots, this cluster is made up of six titles, which are all novels for twelve-year olds written before 2014: *Skellig*, *The Fire-Eaters*, *Heaven Eyes*, *Kit's Wilderness*, *Secret Heart* and *Counting Stars*. Almond's writing style in these books is very similar, which is evident from the distinct cluster they form. This similarity can be due to the intended readership, but can also be explained by the period in which they were published. After all, Almond published these titles one after the other, without bringing out any other books that are included in the corpus of this chapter between *Skellig* in 1998 and *The Fire-Eaters* in 2003.

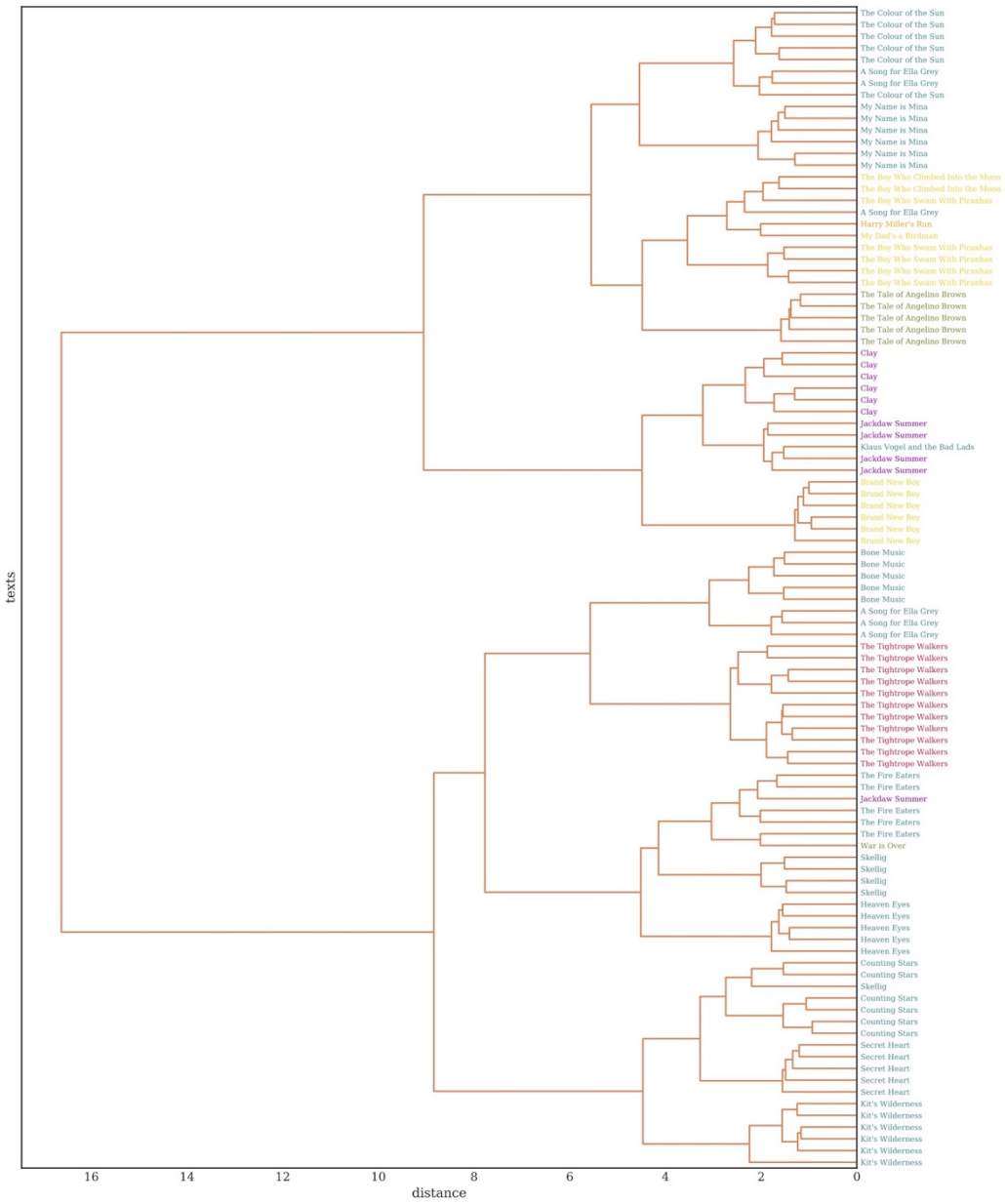


Figure 6.1: Hierarchical cluster tree analysis for David Almond’s novels, colour-coded according to the age of the intended reader.

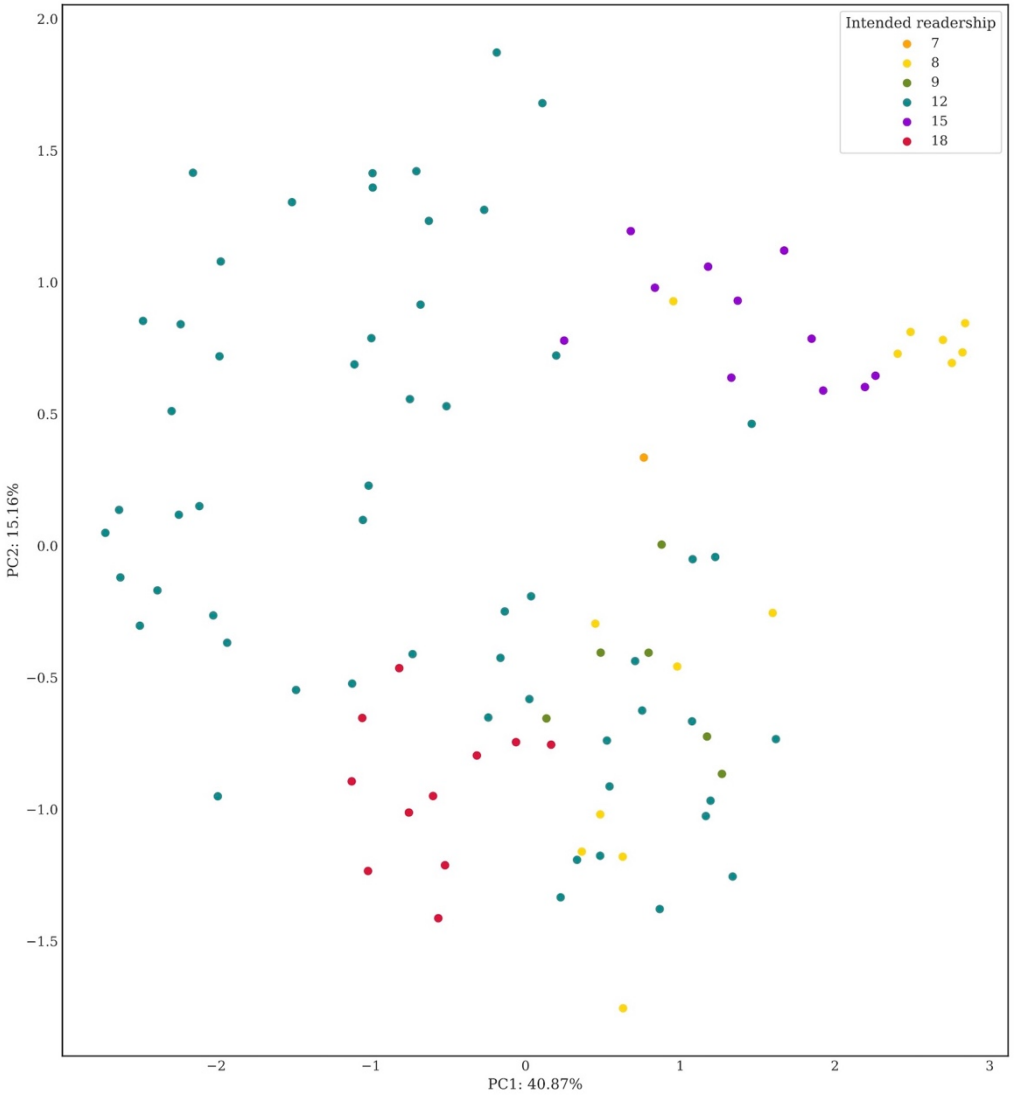


Figure 6.2: Principal component analysis for David Almond's novels, colour-coded according to the age of the intended reader.

The next period is interesting to consider in this light, because from 2009 onwards, Almond alternated books intended for readers of different ages that are all included in the corpus studied for this chapter. With only one exception, his novels for twelve-year olds published between 2009 and 2020 are also stylistically distinct from the books for other ages, as both the HCA and PCA suggest. *Klaus Vogel and the Bad Lads* is the only novel for twelve-year olds that does not fit into any of the previously discussed clusters; in both the HCA and PCA this title is located more closely to the distinct cluster formed by *Clay* and *Jackdaw Summer*, Almond's novels for fifteen-year olds, suggesting that the writing style of this novel is different from the other novels intended for the same age. While it is explicitly marketed for twelve-year olds, as is apparent from Almond's own website, it is written for "struggling, reluctant and dyslexic readers" (Almond, 'Klaus Vogel') and published at Barrington Stoke, which focuses on readable books. The guidelines set by publishers with such objectives might provide an explanation for the deviating writing style, although one might expect that books for struggling teenage readers might be written in a style that resembles books for younger readers rather than for older readers (see also Haverals and Geybels).

I will now consider Almond's books for younger readers. Of his novels for children aged seven to nine, four titles are shown to be stylistically similar in both plots: *Harry Miller's Run*, *The Boy Who Climbed into the Moon*, *The Boy Who Swam with Piranhas* and *The Tale of Angelino Brown*. *Angelino Brown* and *War is Over* are the only two titles written for nine-year-old children and the PCA situates them close together. In Figure 6.1, however, *War is Over* is located in a cluster of books for twelve-year olds and *The Tale of Angelino Brown* is deemed to be stylistically more similar to books for eight-year olds. This suggests that Almond has no distinct writing style when addressing nine-year olds and that these books, more than those for other ages, function as transitional texts between books written for younger and older readers, in this case between children and young adults. Another striking observation that emerges from the plots is the location and distribution of the samples of *Brand New Boy*, one of Almond's most recent novels, published for eight-year-old readers. The six samples that were extracted from this text are clustered together with books for fifteen-year olds, showing little stylistic similarity with other books written for the same age. In short, the stylometric analyses reveal a consistency in style for 12-year-old readers that were published in the period 1998–2003, and stylistic similarity in some books for younger readers. However, various seemingly more random clusters show that Almond does not typically adapt his style for a young audience.

6.6 Speech

After this reflection on style and the age of the intended reader, I now turn to the construction of age on the level of Almond's fictional characters, first by looking at the implicit representation of age through direct speech.[ix] This analysis focuses on vocabulary to gain insight into the similarities and differences between the word choice and topics of interest of speakers who are at various stages in the life course. Characterisation through speech is situated at the midpoint of Nikolajeva's scale from direct to indirect characterisation (*Rhetoric*, 157–58). The starting point of the analysis is the question of whether there is lexical variation across age groups and if so, whether we can see a link between this variation and the age of the intended reader.

To answer these questions, I first extracted all instances of direct speech from the annotated texts and divided them into four files, one for each age group: children, adolescents, adults (ages 20–59) and old adults (ages 60 and up). Stop words, names and words that are only used by one character were removed from these files, as I am particularly interested in words that express topics and that are used by various characters. Next, I used the open-source tool *Scattertext*, developed by Jason Kessler, to visualise how texts belonging to two different age categories (or two different files) differ from each other, based on word frequency.[x] *Scattertext* is an interactive tool, which creates an HTML-file that allows the user to look for specific words in a scatterplot (see appendix B), where each word is represented by a dot. If the user clicks on this dot, the programme gives more information on the words' relative frequency and shows their use in context. I use this tool to find out what words and topics are distinctive for specific age categories. Identifying these topics relies on the researcher's interpretation. This leaves some room for confusion as one word often has different meanings; for example the term 'book' can refer to the noun as well as the verb, and the words 'young' and 'old', which are so obviously significant for analysing age, can refer to people but also objects.

6.6.1 *The Tightrope Walkers*

I begin with the character speech in Almond's adult literature. As this concerns only one novel, *The Tightrope Walkers*, the scatterplots are scarcely populated and no old adult characters are included, as none speak in the novel. Figure 6.3 (see appendix B1 for a larger figure) shows a comparison of child and adolescent speech (on the vertical axis) to speech of adult characters (on the horizontal axis) in *The Tightrope Walkers*. It immediately reveals Almond's distinct use of colloquial language in his writing. Almond draws inspiration from how real people talk in the North of England, where he grew up and where most of his stories are based (*David Almond Interview*). His northern voice is evident in the characters' speech of all ages in *The Tightrope Walkers*. The top left corner of the scatterplot, where terms distinct for children and adolescents are located, features terms including 'bugger', 'tek', 'dunno', 'bliddy', 'nowt' and 'ok', while adults, whose high-frequency words are found in the bottom right area, use 'lads', 'bonny', 'nae' and 'oot'. The scatterplot also reflects the story's main theme of the conflict between freedom and love on the one hand and recklessness and cruelty on the other, which is embodied by the protagonist Dominic. Both young and adult characters talk of darker themes. In the top left corner, which features words distinctive of the younger characters, there are words like 'kill', 'die', 'knife' and 'hell'. If we break these age categories further down we see that 'kill' and 'die' are present almost exclusively in adolescent speech. The speech of adult characters features distinct words related to dark themes, such as 'dead' and 'death', but also its counterpart 'life' together with the term 'love', which has a relative count almost three times higher than in younger characters' speech.

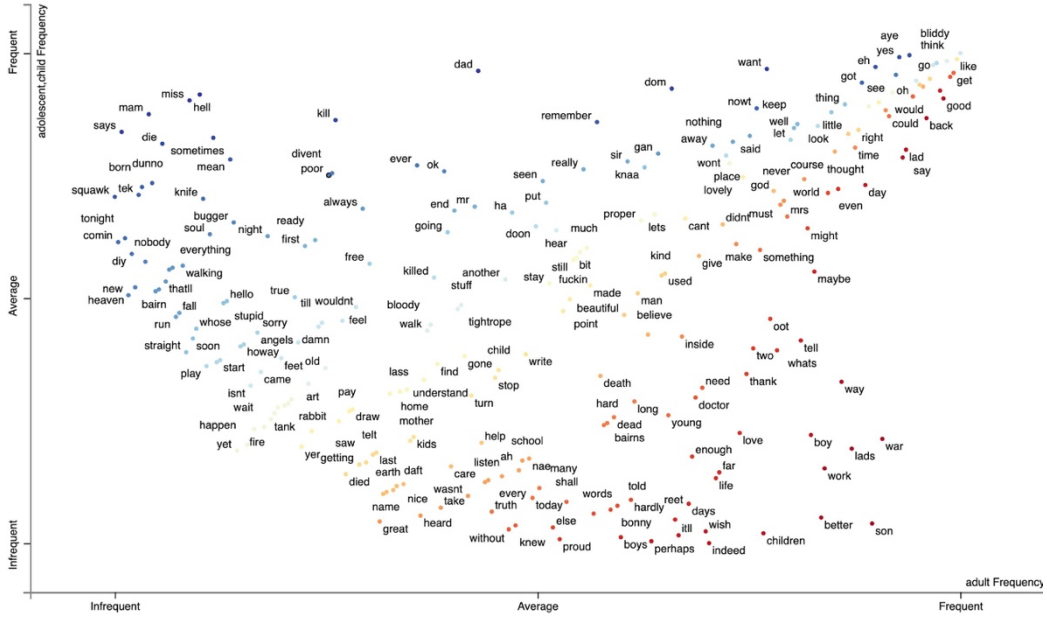


Figure 6.3: Scatterplot comparing the speech of child and adolescent characters to speech of adult characters in David Almond's novel for adults, after removal of stop words and character names.

Also found in opposite areas of Figure 6.3 (appendix B1) are age-related forms of address, suggesting the presence of intergenerational relationships in the novel. In the speech of child and adolescent characters there is a higher frequency of the words 'mam', 'dad', 'miss' and 'mr', pointing toward parental and student/teacher relationships. These words are more distinct to child speech (appendix B2), which can imply that there are more interactions between children

and adults than between adolescents and the latter. The frequency of words relating to childhood in adult speech ('children', 'son', 'boys', 'boy', 'lads' and 'kids') shows the centrality of youth in this novel, even though it is published for adult readers. Other terms with a high frequency in adult speech are 'need', 'care', 'doctor', 'work' and 'war', which match traditional expectations of adult life. As observed by Leander Duthoy, adults in Western culture are expected to have more responsibilities than younger people. This is confirmed in a study by Thomson et al.; when 100 young people from the UK were asked what, to them, constitutes adulthood, two of the most prominent aspects that were named were responsibilities and entering full-time work. In the speech of adult characters in *The Tightrope Walkers*, there are references to responsibilities concerning care, providing for their families by maintaining a job and protecting what is dear to them (although one can argue that waging war is not the best way to do this). Even the term 'school', which can be seen as the counterpart of 'work' as it is one of the few obligations for young people, shows up more often in adult speech, while 'play', associated more with freedom rather than responsibility, is mentioned more by child and adolescent characters.

6.6.2 Young Adult fiction

Many of the same topics from Almond's adult novel return in the scatterplot generated from the speech in his young adult fiction. The thirteen novels in question feature a wide variety of characters, including speech from old adult age (Figure 6.4 and appendix B3). Here too, characters of all ages use colloquial and regional language, but these words are less distinct for the speech of children (with 'nit' and 'bloody' as exceptions) and old adults (with only 'ah' and 'lass') than for adolescents and adults, which both feature a longer list of colloquial terms (appendices B4 and B5). The conflict between death and love in *The Tightrope Walkers* also appears in the scatterplot of Almond's young adult novels (Figure 6.4 and appendix B3). Contrasting words including 'knife', 'murder', 'killed', 'dead', 'war' and 'hate' are distinct to the speech of child and adolescent characters, while 'peace' and 'love' are more characteristic of adult and old adult speech. Upon closer inspection, almost all terms belonging to the darker theme feature most in speech by adolescent characters, except for 'murder' (appendix B4). This term shows up relatively more often in child speech but is used by only two characters from the same novel (*Heaven Eyes*), while the adolescents who talk about murder feature in more than one novel.

The distribution of words referencing other fictional characters in Almonds' YA fiction resembles that in *The Tightrope Walkers* (Figure 6.4 and appendix B3): 'grandpa', 'mum', 'dad' and 'sir' are more distinct to speech of young characters, and adults and old adults use 'son', 'child', 'boys' and 'lad' more often. However, in the Young Adult novels, the young characters also use 'brother' and 'sister', and the adults 'men', 'woman' and 'madam'. This suggests that while intergenerational relationships are still very important in young adult fiction, talking about, or to, peers is more common than in the previous analysis. This is consistent with one of the most common patterns in young adult literature: conflict with parental authority (Trites *Disturbing*, 54) and "the desire to belong to the tribe of [...] peers" (Cockrell, 21; see also Ghesquière et al., 382). If we further break down forms of address by age category (appendix B4), we see that child characters use 'miss', 'mrs' and 'mr' more frequently, while only 'sir' is slightly more distinct to the speech of adolescents. As with Almond's adult novel, this might point towards a higher degree of interaction between children and (old) adults. A similar image emerges when comparing speech of the two adult age groups (appendix B5); adults talk about or address more people of different ages ('son', 'child', 'lad', 'woman', 'madam' and 'baby') than older adults, suggesting that the

older characters in Almond's Young Adult novels are more isolated from intergenerational social networks than other adults.

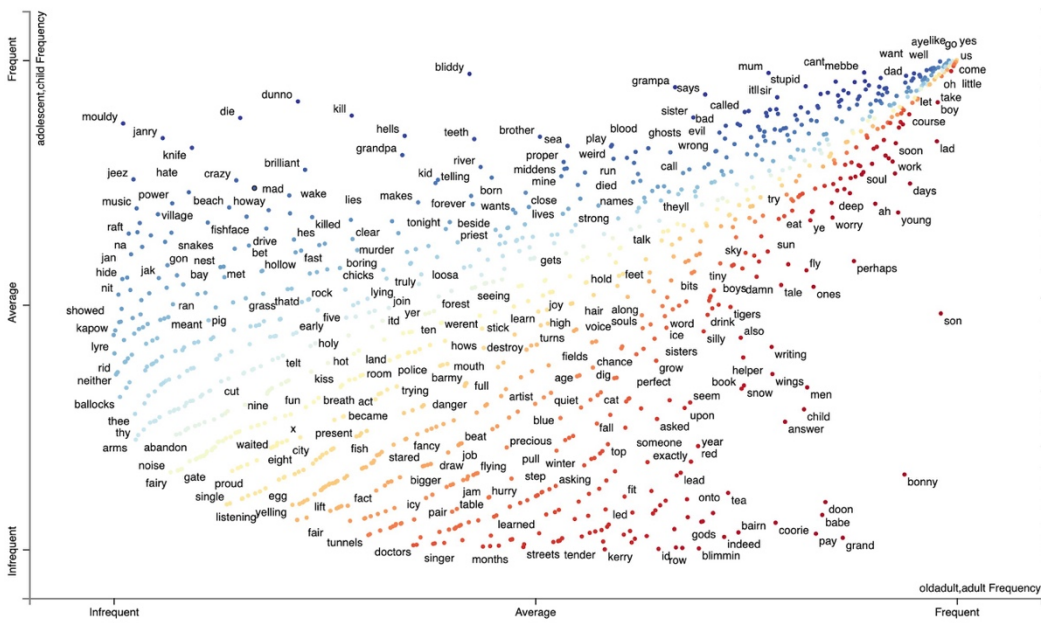


Figure 6.4: Scatterplot comparing speech of child and adolescent characters to speech of adult and old adult characters in David Almond's novels for young adults (aged twelve to fifteen), after removal of stop words and character names.

When revisiting the theme of adult responsibilities, the dichotomy of work versus play as identified in *The Tightrope Walkers* reappears in Almonds' YA fiction (Figure 6.4 and appendix B3). The more detailed graphs present a gradual evolution in responsibilities throughout the life course: 'freedom' is distinct for child speech when compared with adolescent speech, where 'play' and 'playing' show up more. The terms including 'pay', 'work' and 'school' are more used by adults, but not by older adults. This observation is reminiscent of the rather ageist disengagement theory by Elaine Cummings and William Henry, which states that adults give up their responsibilities and withdraw from society when they reach old age. However, when conducting a close reading of Almond's works, various older figures, including Andreas in *Bone Music* and Kit's grandfather in *Kit's Wilderness*, still contribute to their social networks, albeit not professionally.

In addition to the themes identified in *The Tightrope Walkers*, other topics are prominent in the speech of characters in Almond's novels for young adults (Figure 6.4 and appendix B3), including the autobiographical elements of writing, religion and nature that Vanessa Joosen ('Counting Stars') and Emma-Louise Silva highlight. The words 'book', 'books', 'writing' and 'tale' are more distinct for the speech of adult and old adult characters (Figure 6.4 and appendix B3), but feature in the speech of characters of all ages (appendices B4 and B5). However, on closer examination, there are clear differences. For children and old adults, terms associated with stories are much more related to specific characters (the only child character who uses 'writing' is Mina and only two old adult characters use the word 'book'), while they pop up more widely in the speech of adolescents and adults. Less prominent in the speech of younger characters is the theme of religion, which is a clear topic for adults and, more specifically, old adults ('pray', 'god', 'blessed', 'souls' and 'sin' are distinct for their speech). Adolescents do utter the words 'jesus', 'lord', 'hell' and 'heaven' quite often, but the first three are used almost exclusively as exclamations rather

than in the context of a conversation about religion. The theme of nature is distinct in the speech of old adult characters when compared to adults, with words like 'water', 'sea', 'trees', 'fields', 'sky' and 'earth' (appendix B5). This is in line with various classics in children's literature in which an older character shares with a younger character a love for nature (Joosen, 'Second Childhoods')—a trope that Almond also addresses in Young Adult novels such as *Kit's Wilderness* and *Bone Music*.

6.6.3 Books for child readers

Where children's literature traditionally focuses on child characters to facilitate identification and empathy for its intended readership (Nikolajeva, 'Changing Aesthetics'), it is striking that in the seven novels for children that are included in the corpus, more than half of the character speech is spoken by adult characters. Less surprising is the small share of adolescents in this category, with only 90 words in total attributed to characters in that age range. Given that limited word count, I will not go into any differences between child and adolescent characters in Almond's children's books. To a large extent, the same themes that were discussed above surface in Almond's children's books (Figure 6.5 and appendix B6). Once again, colloquial language is used by characters of all ages, with words including 'dunno', 'nowt' and 'mebbe' being distinct to speech of young people and 'ah', 'ye' and 'aha' to that of (old) adult characters. The use of accents and non-standard English is unusual in children's literature, as it may provide a hurdle for young readers and books are also used to teach children standard language. Almond explains that he had difficulty finding agents and publishers due to the combination of his target audience and his northern voice, but that he finds it important to "write about [his] landscape in [his] language" (*January Interview*), also when addressing children.

The terms associated with intergenerational relationships and (adult) responsibilities also show similar trends to those identified in the young adult and adult novels, while terms referring to stories, religion and nature surface less in the direct speech. Children and adolescents address or talk about parents and other adults ('sir', 'ms', 'mr'), but also their peers ('mate', 'pal', 'friend'). The same is true for adult and old adult characters, whose distinct terms include 'brother', 'sister', 'child', 'son' and 'boy'. Different from previous observations is that adults, when compared to old adult characters (appendix B8), use nouns referring to parents ('mam', 'dad') as well as children. This may have several explanations, including adults in children's fiction speaking to children who are not their own (for example: "Does your mam know you're here?") or adults speaking about their own parents. Terms associated with (adult) responsibilities centre mostly around work and school and, like in the young adult novels, all of them are distinct for the speech of adult characters. Work and school may be more prominent in books for children as more words show up in the plot that are associated with them, including 'boss', 'teacher' and 'professor'. In the scatterplot for the children's books' direct speech (Figure 6.5), darker themes surface again, but the more graphic 'murder', 'killed' and 'blood' are rarely found. The theme is also mitigated by the terms 'live' and 'born' that show up in children and adolescent's speech. Death is considered a more typical theme of fiction for adolescents than for younger children (a.o. Talley, 232), which is confirmed here.

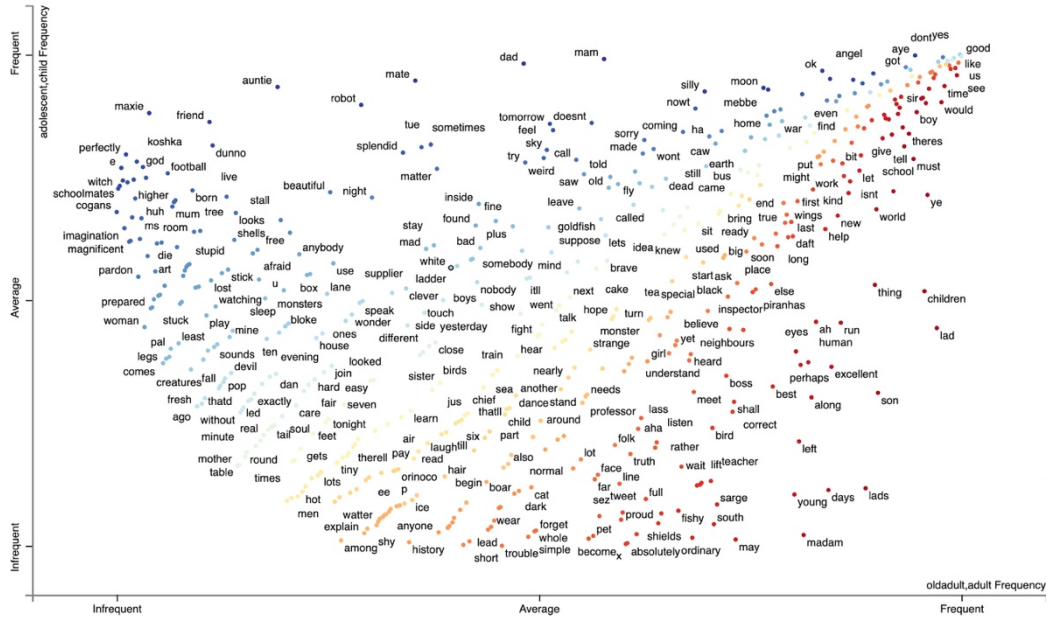


Figure 6.5: Scatterplot comparing speech of child and adolescent characters to speech of adult and old adult characters in David Almond's novels for children (aged seven to nine), after removal of stop words and character names.

Taken together, the analyses above reveal recurring topics as well as some remarkable variation between texts written for different intended ages. Almond's colloquialism shines through in the speech of all his characters. However, in his books for young adults, it is more distinct to adolescent and adult characters. The main topics that are found across the three analyses are intergenerational relationships as well as love and cruelty. In Almond's books for younger readers, peers are commonly referred to, both by young and adult characters, while in his young adult novels, older adults are less often explicitly mentioned or addressed. Violence is most prominent in Almond's young adult novels, where terms associated with cruelty are uttered mostly by adolescents and the topic of love is present in the speech of adult and old adult characters. In *The Tightrope Walkers*, the only adult novel included in the corpus, darker themes are more manifest, mainly in the speech of adolescents and adults. While still present, cruelty is less graphically evoked in character speech in Almond's books for children. There it is mainly present in the speech of young characters and balanced with terms associated with life.

6.7 Action, possession and description

In this final part of my analysis of age in Almond's oeuvre, I discuss explicit character building by looking at the verbs, possessions and adjectives associated with different characters to study their description and actions. According to Svenja Adolphs, these are the two fundamental methods of characterisation: "The way in which a character is depicted in a story is often expressed in the verbs and adverbs that illustrate his or her actions and in the adjectives that are used to describe the person" (67). To study verbs, adjectives and possessions in the annotated files of Almond's work, we developed a so-called 'parser' that is able to recognise the grammatical relationships in a sentence and match verbs, possessions and adjectives with the characters they refer to.[xi] Combining the resulting data with the corresponding character lists (see above) allows the parser to extract relevant information of characters in different age groups.[xii] To keep the discussion concise and relevant to a study of age, my analysis is based on the 40 most fre-

quently used terms for each category (see appendix C).[xiii] The discussion focuses on general trends rather than listing all terms extracted from the text.

6.7.1 Actions

First, the focus of the characters' actions lies on verbal and non-verbal expression, movement and leisurely activities. Across Almond's oeuvre, child characters verbally express themselves mainly by whispering, giggling, yelling and laughing. While their inquisitive nature is reflected in the novels for children and young adults, where 'ask' is a high-frequency verb, the adult novel *The Tightrope Walkers* paints a more stereotypical picture; it is the only text to include 'cry' in the 40 most common verbs for children. Overall, adolescent characters express themselves similar to characters from other age groups, but it is noteworthy that they are the only ones to frequently 'mutter' and 'gasp'. It is striking how often adolescents 'shrug' across the corpus (adults do so too, but far less often). Not surprisingly, as they make up the largest portion of the characters featured in *The Tightrope Walkers*, adult characters show a diverse array of ways to express themselves verbally and non-verbally in this adult novel: adults 'laugh', 'tell', 'ask', 'whisper', 'smile', 'sigh', 'yell', 'speak', 'grin', 'call', 'cough' and 'murmur'. Contrastingly, in books for young adults and children, there is less variety in the ways that adult characters communicate; there, they primarily 'laugh' and 'smile' in the books for the youngest readers. Featured mainly in books for young adults and children, some of the verbs indicating how older characters express themselves, such as 'murmur' and 'sigh', tend towards a stereotypical image of unhappy older people, but the opposite is also shown with 'laugh' and 'smile'.

A second trend in the 40 most common verbs associated with characters of all ages pertains to leisure activities and creativity. The novels for adults and young adults portray children as creative beings who 'draw', 'dream', 'read', 'write', 'imagine' and 'remember'. Adolescents engage in similar activities throughout Almond's oeuvre. Apart from the substitution of 'paint' for 'draw', the most common leisure activities of children and adolescents are identical. By contrast, the material interaction of adults is rather abstract, as suggested by the verbs 'touch' (in adult and young adult literature), 'open' (in novels for young adults and children) and 'unfold' (in children's books). In Almond's novel for adults, the image of childhood is complicated by the verb 'kill'. As discussed in the previous analysis, this is most likely due to the theme of cruelty present in the novel. Parallel to younger characters, old adults are more associated with leisure activities; they 'write', 'remember' and 'dig' in young adult fiction and 'drink' and 'drive' in children's books, although hopefully not simultaneously.

Notable differences between characters of various ages are also found in terms of movement; children and adolescents 'run' (or 'outrun') significantly more than adult and old characters. This activity occurs more frequently in novels for children and adults than in young adult fiction. Contrastingly, adults (in adult and young adult novels) as well as old adult characters (in young adult and children's literature) often 'lean'. Other verbs pertaining to movement associated with old adults, who do not feature in the adult novel, are largely stationary in nature; rather than running and walking, they 'stand', 'sit' and 'turn'.

6.7.2 Possessions

As explained above, 'possessions' are here understood in a grammatical sense. The parser recognises genitive forms and possessive pronouns and can thus link attributes to the corresponding characters. Because there are few adolescent characters in Almond's books for children, the list of the 40 most frequently occurring possessions of this age group contains only hapax legomena, or terms which occur only once, and will therefore not be discussed. Most of the extracted possessions for other age groups relate to the body, community versus home, and work. The largest category in the 40 most common possessions of characters of all ages pertains to the human body. Most body parts are found in all age groups, but some are characteristic of a particular group. Child characters in books for children and adolescents frequently possess a 'voice', but not those in the adult novel. This is in line with the observation in the analysis of character speech that children hardly speak in *The Tightrope Walkers*. In that novel, 'blood' and 'brain' are distinctive of the adolescent body. The former may hint at the aggressive nature of adolescent characters, an interpretation reinforced by the high occurrence of 'knife' in both the young adult and adult books and 'mask' in the latter.

More overlap can be found between the analysis of direct speech presented above and characters' possessions; many terms bear evidence of intergenerational relationships or, more appropriately here, the communities surrounding characters. Joosen ('Counting Stars') expands on the element of community in *Counting Stars*; my large-scale analysis makes clear that family and friends play an important role throughout his oeuvre. The members of the community of child characters most frequently referred to across all of the author's works are parents and friends. In young adult fiction, 'siblings' are also frequently mentioned, while 'uncle' and the more general 'family' feature in the 40 most common possessions of child characters in children's books. Keeping the observations from the previous section in mind, it is worth noting that any references to grandparents are missing. However, this does not necessarily contradict the intergenerational bond between children and their grandparents but suggests that grandparents are mostly talked about within the family, without the possessive pronoun (e.g., 'granddad' instead of 'my granddad'). In the adult novel *The Tightrope Walkers*, adolescents are also represented as being parents themselves, as they have a 'baby'. References to children also show up amongst adult possessions, in all categories of intended readers, and are predominantly male: 'son', 'boy' and 'baby'. The community of adult characters is considerably less extensive and comprises almost exclusively family; the only references to adults' friends through possessive forms are made in the adult novel. The community of older adult characters referred to with possessive forms consists exclusively of family.

Complementing the theme of community are terms associated with private space. In books for adults and adolescents, there are frequent references to the 'bed' of a child character and in young adult and children's literature there is talk about their own 'room'. In the YA novels, 'room' is featured in the list of common possessions of adolescents and their 'house' is frequently referred to in adult literature. For adult characters, the terms 'room', 'house' and 'apartment' point towards interactions with private spaces. Those of old adults are not limited to indoor spaces; together with 'apartment' and 'caravan', 'garden' shows up in the 40 most common possessions of old adults in books for children. In young adult literature, the space is again more restricted and older characters only have a 'window' through which they can look at the outdoors from their 'bed' and 'room'.

What is more prominent for (old) adult characters, at least in books for adults and children, are material possessions, mainly pertaining to outward appearance ('jacket', 'bag', 'collar', 'glasses' and 'cigarette'), and objects associated with professions ('desk', 'office', 'uniform', 'work' and 'colleague'). 'Work' is also included in the 40 most common possessions of child and adolescent characters in Almond's novel for adults but might refer to tasks assigned to younger people at school rather than to a professional job. While present in the lists of characters of all ages, the theme of responsibilities discussed above is more prominent in the list for adults. The allusion to disengagement theory in the discussion of speech of old adults, where the topics of adult responsibility and social networks were less present, is nuanced in the study of their belongings. The theme of responsibilities, especially centred around work life, remains present; in young adult fiction, these characters have a 'desk' and in children's literature the older generation is frequently attributed 'ambition', a 'job' and a 'report'. Similar to adults, terms associated with physical appearance are also featured in the most common possessions of old adults in young adult fiction; they are the only characters to frequently have a 'beard' as well as wear a 'helmet', 'jacket' and 'suit'.

6.7.3 Adjectives

Finally, the discussion of adjectives in Almond's works focuses on characters' physical size, character traits and physical appearance. I will start with adjectives describing the characters' bodies. Many of the most common adjectives in Almond's work validate the stereotypical image of the small and vulnerable child. However, there is a surprising difference in connotation between terms found in books for different intended readers. A central theme in YA fiction is the transition from childhood to adulthood, and one might expect that adolescent characters might be eager to distance themselves from the younger age group. In Almond's YA novels, however, a more caring stance emerges, with child characters being described as 'delicate' and 'precious'. These endearing words stand in stark contrast to how child characters are portrayed in narratives for children as 'ickle' and 'scrawny'. Adolescents in *The Tightrope Walkers* are repeatedly called 'little', 'sweet' and 'silly'. The young adult books paint a diverse picture. In terms of physical size, adolescent characters are both 'little' and 'big' but are more often than not 'slender'.[xiv] The physical appearance of adult characters, for its part, is described in varied ways in children's books and thus seems to be an important aspect of their characterisation. Besides the highly frequent 'big' and 'little', their stature is referred to as 'slender', 'burly' and 'chubby'.

A predictable image of child characters emerges from adjectives related to character traits. In young adult fiction, the recurrent adjectives to describe children suggest balanced traits: 'ignorant' and 'scared' as well as 'strong' and 'lucky'. In children's books, child characters are mainly described in a positive light with an emphasis on their intellect and all-around good nature, with terms including 'clever', 'bright', 'gifted', 'perfect', 'fantastic' and 'brave'. The adolescent characters in young adult fiction are regularly described as 'dreamy', 'clever', 'brilliant', 'lucky', 'sweet' and 'brave', but they are also frequently 'troubled', 'shy', 'bad', 'daft' and 'silly'. The theme of societal alienation that is often present in young adult fiction and adolescent characters is also apparent from the most common adjectives in this category. More than any other age group, adolescents are described as 'strange', 'weird' and 'different'. In terms of character traits, the representation of adult characters in young adult novels is also balanced. Adults are deemed 'extraordinary', 'nice' and 'great', but also 'crazy', 'useless', 'awful', 'boring' and 'reckless'. The latter is contradicted in Almond's books for children, where adults are pre-

dominantly described in positive terms as being ‘sensible’, ‘great’, ‘nice’ and ‘clever’. Their only flaws are rather lovably described as ‘silly’, ‘barmy’ and ‘daft’. Adjectives describing old adult characters that are not shared with other age groups are too limited to identify trends with any certainty. Very superficially, a more positive image emerges from the terms ‘sweet’ and ‘nice’ in children’s literature than from ‘filthy’ in young adult fiction, but all of these words occur only twice.

6.8 Conclusion

Digital tools can reveal implicit age norms that stay under the radar and that can, according to Stephens (*Language*, 10), be all the more effective in influencing readers because they are not explicitly addressed and put up for scrutiny. The stylometric analysis of Almond’s work shows that overall, there is a stylistic similarity between novels intended for readers of the same age. However, there are exceptions; some titles (e.g., *Klaus Vogel and the Bad Lads* and *War is Over*) are not included in the cluster of texts for the age of their intended readership, but are grouped together with others. Books for twelve-year olds and fifteen-year olds seem to be stylistically most distinct from others, even if some samples from these novels are still clustered with books from other age ranges. Almond’s writing style is not distinct for each age range and does not show a gradual change from the youngest to oldest readership, a trend that can be identified in some authors’ oeuvres but that is overall rare (see Haverals et al.).

Turning to the age of fictional characters, the analyses of speech and possessions show that child characters in Almond’s children’s books typically have an extensive community around them consisting of friends, family and other adults. In young adult fiction, this group further includes siblings and grandparents. Thus, although intergenerational relationships feature in all books, they are particularly extensive in books for younger readers. This age norm emerges both from the close reading of his work and the computational analyses. The characters in Almond’s work who lack this community, usually adolescent boys, often display strange and aggressive behaviour that is also captured in the vocabulary related to them. In books for all ages, but in his adult novel in particular, Almond’s adolescent characters are associated with the theme of death. In his young adult and adult books, the adjectives evoke the image of adolescents who are rebellious and alienated from society. These similarities between adult and young adult literature are acknowledged by the author himself: “even when I write for adults, there is usually some kind of sense of youth, or a young protagonist” (*Interview David Almond*). Although the theme of alienation is strongly present, there is a community surrounding (some) adolescent characters in all books. Their size and character traits are evoked as balanced, with positive and negative aspects. The descriptions reveal that Almond’s adolescent characters are often evoked as skinny, lean figures. From the analyses of speech and actions, adolescents, parallel to child characters, are associated with creativity and, consequently, with the leisure time needed to express this creativity.

Almond’s books present a diverse picture when it comes to childhood. In books for children, the negatively connotated description of their size is contrasted with positive attributes, specifically their intellect. Conversely, in his YA novels, children’s small size is described in endearing terms, while their personality comes across as more balanced. Based on the character’s vocabulary and actions, the freedom and creativity traditionally associated with children are most evident in Almond’s adult literature and young adult fiction.

For adult characters, creativity and leisure make way for responsibilities. In books for readers of all ages, there is an association between adulthood and professional life in the analyses of speech and possessions. In the books for children and adults, there is also the responsibility of care, which goes hand in hand with the community surrounding them, which consists mainly of children and is particularly evident in the analysis of speech in all the books. In children's novels, some adults have more varied relationships, including siblings and possibly their own parents. There, the digital analyses suggest that only in the adult books do friends of adults appear frequently. There they also express themselves in varied ways, whereas the analysis of actions shows that in children's books they mainly 'laugh' and 'smile'. In children's books, adults' personalities are largely described in a positive way (even if these books also feature some adult villains), while a more balanced picture emerges from the books for young adults. This might be explained by the demands associated with young people; while children are more dependent on parental figures for basic needs such as care, adolescents are more independent and defiant towards them.

Older adult characters are less common in Almond's books, and absent from his adult book. For the digital analysis of his books for children and young adults a similar image emerges. The speech and possessions of both categories suggest that the community around old people is considerably smaller. The stereotypical image of the disengaging older adult is reinforced by a greater emphasis on private spaces. Then again, the responsibility of work is mentioned for these characters, albeit much less prominently than for adults. In both children's books and young adult books, the decline narrative is evoked when old people are described with terms that express helplessness, fragility and static movement. In young adult novels, this concept is brought to the next level as the verb 'die' emerges in the analysis of actions.

The analyses in this chapter form an important contribution to the existing body of research on David Almond as a crosswriter. Using computational methods allowed me to discuss more titles than any previous studies and study the works together instead of treating them separately. This made it possible to compare the writing style of texts, discover trends in specific aspects (e.g., attributes of fictional characters of specific ages) and compare their speech. A limitation that this method has, however, is that some titles had to be excluded from the corpus because they did not contain enough text or were written in an idiosyncratic language. This complicates the argument that digital analyses can provide an objective and data-consistent analysis. Additionally, relying on lists of most common words to identify predominant themes in Almond's work overlooks potentially significant topics that are described using a diverse vocabulary, where each term is used only sparingly.

The benefit of studying many texts at once is that I could identify some variety in texts intended for different ages; while the theme of death, intergenerational relationships and the child's need for protection recur in all categories, they are addressed differently depending on the age of the reader. Violence and death are presented more graphically in books for young adults and adults. Adults are placed in a more positive light in books for children than in those for young adults. Child characters are described as vulnerable in all books, but this characteristic is compensated by more agentic features in children's books. Many of the trends identified in this chapter can serve as hypotheses to be tested with close reading, which remains a valuable tool to study the representation of age in fiction and might enrich or nuance findings that come out of a more distant approach as employed in this chapter.

[i] Stylometric analyses were already being conducted before the use of computational methods. The most influential early study making use of stylometric analyses was a case in disputed authorship, undertaken by Mosteller and Wallace in 1964, who moved from the analysis of low-frequency features to high-frequency features in a study of style. Since then, this has been the most common method in stylometry, also used by Celia Catlett Anderson in her doctoral dissertation in 1984, where she went beyond the research of reading specialists, who examined vocabulary as well as the structure and length of sentences for prescriptive purposes, and moved towards a general analysis of the characteristics of children's literature (39).

[ii] See Cadden, 'Narrative Theory and Children's Literature' for a brief overview of studies of character in children's literature.

[iii] Furthermore, the image that a reader creates of characters through their speech is often affected by dialogue tags which express the agency of the narrator (Nikolajeva 'Changing Aesthetics', 446). In this chapter, direct speech is separated from narration, removing all context from the discussion and thus moving speech more towards the side of direct characterisation. The speech tags are taken into account as part of the digital analysis of verbs that are linked to the characters.

[iv] An exception is made for Counting Stars, because the same characters appear throughout the publication, which can be read as a single narrative. Primers and picture books are mainly directed to first readers and thus use short sentences and a limited vocabulary when compared to other children's books. Pertaining to the second set of analyses, those on the construction of the age of fictional characters, an important part of character building in picture books happens visually, through the illustrations included in the work. Because the digital analyses included in this chapter are exclusively text-based, a lot of information on the representation of characters and more specifically their age, would be lost. This is also the case with graphic novels and plays.

[v] To facilitate the task of the human annotator, a self-developed Python script is used to automatically insert tags as the texts go through a pre-annotation process.

[vi] The field of digital humanities provides a tool to classify named entities in a text, including person names, called Named Entity Recognition (NER). However, NER was not used in the annotation process of Almond's works for two reasons. First, none of the existing NER models classify pronouns, which means that many character references would not be identified. Second, automated character recognition does not consider the element of age which is central to the question of characterisation in this chapter. Even if there were a model that could correctly classify pronouns, it would be an insurmountable task to automatically trace changes in a character's age throughout a text. To that effect, tags are added to all pronouns during the pre-annotation process, while proper names and nouns referring to characters are manually tagged by the human annotator, who also attributes the relevant character ID to each tag.

[vii] The Python script used for the stylometric analysis in this section was written in large part by Wouter Haverals for the article 'A Style for Every Age: A Stylometric Inquiry into Crosswriters for Children, Adolescents and Adults', co-authored by Vanessa Joosen and myself, to which I refer for technical details. I made the following updates to the script: First, rather than a full manual pronoun culling, I make use of the syntactic parser software spaCy to automatically remove unwanted words. Second, following Maciej Eder's advice, the remaining words are randomised before sampling the texts to avoid clustering due to the disparity in the distribution of vocabulary between narrative and dialogue.

[viii] The HCA (Figure 6.1) is generated using scipy's *cluster.hierarchy.dendrogram()* function. The following settings were used to generate the HCA and PCA: features n=54, chunk size=5000, n-gram size range=(1,1), min_df=0.1, max_df=1.0, vectoriser=TF-IDF, distance metric=cosine similarity.

[ix] The discussion in this section builds on preliminary research carried out by intern Jo Debruyn who worked on the CAFYR research project in 2021-2022. The author is indebted to her for the insights she brought to this analysis.

[x] See Joosen ('Research in Action') and Haverals and Joosen for the implementation of this tool to the oeuvre of Bart Moeyaert and Guus Kuijer, respectively.

[xi] The Python script was developed in collaboration with TEXTUA, the Text Mining Centre of the University of Antwerp and is available in open access via the following link: http://textua.uantwerpen.be/app/cafyr. It uses the open-source software *spaCy* (Honnibal and Montani) to parse the syntactic dependencies in sentences.

[xii] A prototype of this parser was used by Joosen in 'Research in Action: Constructing Age For Young Readers' (2021), published in *International Research in Children's Literature* to study adjectives in Philip Pullman's *La Belle*

Sauvage. See this article for a brief introduction to previous studies conducting analysis with other implementations of automatic syntactic dependency parsers.

[xiii] In the discussion, we omit words that appear in all or almost all lists as they do not contribute to a comparison of the representation of different age groups. Likewise, terms that only occur once for any given age range are also excluded, as these may be heavily dependent on specific characters. Adjectives that are discarded include ‘good’, ‘lovely’, ‘poor’ and indications of age, as characters of all ages are described as being both ‘old’ and ‘young’, the only exception being old adult characters in young adult fiction, who are exclusively described as ‘old’ and ‘ancient’.

[xiv] Similar to the possessions of adolescents in Almond’s children’s books, there are no adjectives that are used to describe characters in this age group more than once and they are thus discarded from the discussion.

Chapter 7 Shuffling Softly, Sighing Deeply

A Digital Inquiry into Representations of Older Men and Women in Literature for Different Ages[6]

Abstract – When gender is brought into concerns about older people, the emphasis often lies on stereotypes connected to older women, and few comparative studies have been conducted pertaining to the representation of the intersection between older age and gender in fiction. This article argues that not only children's literature, traditionally considered to be a carrier of ideology, plays a large part in the target readership's age socialisation, but so do young adult and adult fiction. In a large corpus of 41 Dutch books written for different ages, the representation of older men and women is studied through the verbs, grammatical possessions and adjectives associated with the relevant fictional characters, which were extracted from the texts through the computational method of dependency parsing. Older adult characters featured most frequently in fiction for adults, where, more so than in the books for younger readers, they are depicted as being prone to illness, experiencing the effects of a deteriorating body and having a limited social network. In the books for children, little to no association between older adulthood and mortality was found in the data. Ageist stereotypes pertaining to both genders were found throughout the corpus. In terms of characterisation, male older adults are associated more with physicality, including matters of illness and mobility, while character traits and emotions show up in a more varied manner in connection to female older characters.

Context – Chapter seven focuses on characterisation in the Dutch-language part of the CAFYR corpus to study how the construction of age in the text is influenced by the age of the intended reader of that text (RQ 1.2). It also goes further into RQ 2.5 as it uses the syntactic parser introduced in chapter six; words associated with characters belonging to a specific age group, namely older adult characters, were extracted from 98 annotated texts. The article on which this chapter is based is the first study to look at the construction of older age in Dutch children's literature on a large scale. This chapter takes the study of characterisation a step further by looking at the intersection of age and gender, a trait that was tracked for each character during annotation. Due to the narrow age range of the characters that are analysed, only 41 of those texts, written by six different authors for a wide range of readers, yielded relevant data. The benefit of a smaller data set as well as reporting on results of a single type of analysis is that this chapter goes more in-depth than some of the others.

[6] This chapter is based on: Geybels, Lindsey. 'Shuffling Softly, Sighing Deeply; A Digital Inquiry into Representations of Older Men and Women in Literature for Different Ages'. *Social Sciences*, vol. 12, no. 112, 2023.

7.1 Introduction

Recently, stereotypes connected to older women have been widely addressed in different media, from Jessica Bennett's "I am (an Older) Woman. Hear Me Roar" in the *New York Times* to the categories such as "the coastal granny" and "Madonna" that Lieve Van De Velde identifies in *De Standaard* and Michele Hanson's article in *The Guardian* that considers older women as the "glue that holds society together". Older men are also subject to stereotyping, for example, when they are presented as wise mentors or grumpy seniors. However, these images are still more naturalised as their voices are less represented in the mass media. Furthermore, when gender is brought into concerns about older people, the emphasis often lies on "the growing numbers of older women", which according to Jeff Hearn, holds true not only in political settings and popular media, but also in various fields of literature (98). His plea that "the sociology of aging [...] should be fully, not partially, gendered" (ibid.) has evidently not entirely found a reflection in the mass media since he voiced it in 1995, as men are hardly seen berating the ageist representations they undergo based on the combination of their age and gender. By integrating age studies and children's literature studies with the digital humanities, which applies digital tools to study the disciplines of the humanities, this article will gauge how older men and women are portrayed in fiction. Contemporary age studies depart from the premise that age is a social construct as much as gender; as the influential age critic Margaret Morganroth Gullette puts it, "human beings are aged by culture" (Gullette *Aged by Culture*, 12).

Taking note of Hearn's plea to study the relationship between gender and age more completely, the analysis will compare descriptions of older female and older male characters. Moreover, it will consider the differences between books intended for readers of different ages to obtain a better understanding of how the representation of older adults in children's literature compares to adult literature. Children's literature is one of the cultural discourses in which this construction of age takes place (Joosen *Adulthood*, 9); it carries ideologies and reflects, as well as helps to create, societal values, perpetuating descriptions of older characters and launching ideas about how they look, think, behave, and act. Children's books thereby contribute to young readers' age socialisation (Stephens *Language*, Hollindale *Ideology*). This process is not completed at the end of childhood, however, as adolescents remain susceptible to negative images of older adulthood. Gullette states that "ageism may worsen for adolescents, perhaps from overhearing the sotto voce complaints or jokey-fearful allusions of their parents to 'aging'" (Gullette *Ending Ageism*, 57). While Gullette refers to real-life interactions with adults, her observations open the door to the hypothesis that young adult literature continues to contribute to the socialisation of its readers with regards to their perception of (older) age. In contrast to books for young readers, adult novels are targeted at a broad range of life stages, from early adulthood to people in deep old age. Thus, the representation of older adult characters in those books is consumed by people of, or close to, that same age. The representation of older age then becomes no longer merely a matter of socialisation, but also one of self-reflection.

This article delves into the question of how adult authors construct senescence in children's books, young adult novels and books for adults, while keeping in mind the intersection of age and gender. There are several interests in studying older age in fiction for different ages. First, as the average age of the worldwide population is rising, it is important that not only children, but also adolescents and adults, are "aware of age ideology" (Joosen 'Not Your', 182) so that they can also be critical of ageism. Second, Sandra McGuire argues that positive images of ageing are

important for today's children and how they age. While I agree with her statement that "the foundation for positive aging should begin during youth" (2605), I believe it is important not to limit the study of the representation of older age solely to children's literature as the process of developing imaginaries for positive aging continues throughout all life stages leading up to it. While holding a negative image of older age may seem beneficial to younger adults in a more immediate manner, such as paying lower insurance rates, in the long run it proves to be detrimental to their mental and physical health (Levy). To put it another way: today's children, adolescents and adults are influenced by their perception of older age not only in their interactions with older people in their daily lives, but also in the way they look towards their own futures, as they grow towards older age. Third, because of the points raised above, a better understanding of the portrayal of older adulthood in fiction is a necessary step towards equality based on age, whereas, at present, ageism, inequity and social injustice still abound.

In compiling a corpus of Dutch and Flemish fiction, I borrow the reasoning from two publications by children's literature scholar Vanessa Joosen. First, Dutch literature has presented "a preoccupation with broadening the range of age narratives" over the last few decades (Joosen 'Aging', 4). Second, Dutch children's books "display an original and progressive take on old age" in several ways, including "by highlighting the diversity in old age" (Joosen 'Not Your', 182). Rather than selecting literature written for readers of different ages by different authors, the texts in the corpus are all written by crosswriters. Bibi Dumon Tak, Ed Franck, Guus Kuijer, Bart Moeyaert, Hilde Vandermeeren and Joke van Leeuwen have all written novels for at least two categories of readers based on age (e.g., children's literature and adult literature, or children's literature and Young Adult literature). Celia Catlett Anderson argues that comparing passages by authors who address a different readership in different works is the best method to compare children's and adult literature as it minimises the differences according to the individual authors (40). This article answers the following three research questions by studying the representation of older age in their work:

1. How do representations of older adulthood differ according to gender?
2. How do representations of older adulthood differ according to the age of the intended reader?
3. Can we find evidence of Dutch progressiveness by using a distant reading method for characterisation?

This last question serves as an evaluation of the unique methodology this article employs to study older age on a large scale. The representation of fictional characters is studied in a quantitative manner, which relies on different types of descriptions based on the relevant adjectives, verbs and possessions. Studying these word types to better understand the representation of fictional characters is discussed by both the linguist Svenja Adolphs and the leading children's literature researcher Maria Nikolajeva. The former claims that there are two fundamental methods of characterisation; namely, "the verbs and adverbs that illustrate [a character's] actions and in the adjectives that are used to describe the person" (Adolphs, 67). In Nikolajeva's narratological study *The Rhetoric of Character* (2002), she places different methods of characterisation on a scale from direct to indirect methods, from description and narration to implicit characterisation through a number of different constructs. In between these two extremes are the actions a character performs, the events they become embroiled in, their speech and their internal perceptions (Nikolajeva *Rhetoric*, 157–58). In this paper, adjectives represent the direct

characterisation on one end of that continuum, verbs represent the actions taken and possessives are studied as a type of implicit characterisation. Combining these data that can be extracted from a large number of texts through the use of digital tools with insights from the field of age studies can improve our understanding of the construction of older age and how this life stage is represented in books for readers of different ages.

Since the last quarter of the twentieth century, several studies have been published that trace and criticise ageism and stereotypes in books for young people. However, some of these, including Marilyn Apseloff ('Grandparents') and Hilary Crew, have only studied specific facets of older age, such as the representation of grandparent figures and daughter-mother narratives, respectively. Other studies, including Sylvia Henneberg, situate themselves on the intersection of older age and gender, but often investigate only one gender, most often older female characters (see also Chivers, and Woodward *Figuring Age*). Henneberg concludes that classic children's stories perpetuate ageism and sexism. Since the mid-2010s, a higher awareness around adulthood and its construction in children's literature has emerged. The first work to focus specifically on (older) adulthood as a life stage, as opposed to roles such as motherhood, fatherhood and grandparenthood, is Joosen's *Adulthood in Children's Literature*, in which she argues that age studies and children's literature studies have many insights to offer each other, as they both acknowledge that age is a cultural construct as much as it is a biological given (9). In this work, Joosen discusses four ageist tropes: the decline narrative, the infantilised senior, the disregard of the old body and the wise old mentor. The latter is an example of an ageist stereotype as a result of the inclusion of positive images of older adulthood. Age scholars argue that reproducing these types of images disregards the diversity present in older age.

While borrowing insights from the previously mentioned small-scaled studies, this article also builds upon the tradition of large-scale studies into the representation of older age in children's books. From the very beginning of the studies mentioned above, there was an interest in seeking trends in a wide range of texts. In one of the first studies into older age in children's literature, Phyllis Barnum conducted a close-reading analysis of 100 books for children from preschool through to grade three in the United States of America (USA). She found that "the aged play an insignificant role in young children's literature" (303) and that, when including gender into the analysis, older women are even less represented than older men when compared to the real-life distribution of older adults. Furthermore, older adult characters are portrayed as being excluded from society, as well as less healthy and less self-reliant than other adults (ibid., 304). In 1978, Edward Ansello published a paper entitled "Ageism—The Subtle Stereotype", based on research of a corpus of 656 children's books. Older adult characters featured in only 16% of these titles. Ansello studied the representation of these characters through close reading and found mainly stereotypical characterisations. The same conclusion was reached by Linda Janelli's research into depictions of grandparents in 73 books for children in 1988.

In the last decade of the twentieth century, the stereotypical image of grandparents that researchers identified shifted. Beland and Mills studied 64 children's books published between 1985 and 2000 and found that these older adult characters are often characterised as, among others, independent and happy, creating an overall positive portrayal. Ng et al.'s large-scale study of ageism moved towards a digital methodology, in which they identified an attitude opposite to that of Beland and Mills' study. From collocations, words that co-occur frequently in a text, in a corpus of historical American-English of the past 200 years, Ng et al. found that age stereotypes

have gradually become more negative, with a turning point in the 1880s, where they found that stereotypes switched from being positive to negative. While collocation is an accessible tool for extracting associated terms from texts, it has certain limits. For example, Ng et al. used a list of synonyms for the term 'elderly' as the starting point of their analysis, which potentially discounts characters who are not constantly referred to as elderly, senior or older. By using an annotated corpus as the data set, as well as more advanced techniques from the field of digital humanities, this article places itself firmly at the forefront of the search for new methods for studying older age in fiction.

7.2 Materials and methods

The analyses in this article rely on tools from the digital humanities for two reasons. First, they make it feasible to examine specific aspects of a large corpus in an efficient way and to study different avenues of research in a relatively short time, bypassing the need to reread a large number of texts. A second advantage is that they can recognise patterns that are often less obvious in a traditional close reading, such as information about word frequency and the co-occurrence of words. The corpus analyzed below constitutes 41 contemporary books, published between 1975 and 2021 and written by Dutch and Flemish crosswriters. To study characterisation, all of the texts were digitised and an additional layer of information pertaining to the fictional characters was added in a process of textual annotation. First, the books were either obtained in digital versions or physical copies were scanned and converted into editable text documents using Optical Character Recognition (OCR) software. Appendix D provides an overview of the works, along with the age of the intended reader for each, according to which the texts are categorised in the analyses in the next section. With only a few exceptions, most of the books in the corpus are recorded in the Centraal Bestand Kinderboeken (CBK), a Dutch database that includes information on the intended readers' age. For titles not included in the catalogue, the table explicitly states the source of the reader's age information; it was either taken from the physical book or the author's website.

Second, annotations were added to various parts of the digitised texts. By adding tags to pronouns, proper names and nouns, all references to characters were disambiguated. These tags are formatted by using Extensive Mark-up Language (XML) in accordance with guidelines set up by the Text Encoding Initiative, a widely used standard in digital text analysis. The following fragment from Joke van Leeuwen's *Alles Nieuw* (2008) illustrates the texts after annotation:

> This house is a hundred years old, says the <rs ref="ada80s">landlady</rs>, who <rs ref="laraearly">I</rs> think is at least eighty. All predate <rs ref="laraearly">my</rs> time. This is <rs ref="laraearly">my</rs> time. <rs ref="ada80s">She</rs> lives downstairs, <rs ref="laraearly">I</rs> have to pay <rs ref="ada80s">her</rs> in cash every month, <rs ref="ada80s">she</rs> jots this down in a notebook, peering at the tip of <rs ref="ada80s">her</rs> pen, while <rs ref="laraearly">I</rs> sit looking at <rs ref="ada80s">her</rs> hands, at the map of veins and dark spots that could represent rivers and villages.[i]

Each reference to a character is enclosed by a so-called 'rs-tag', which is attributed with an id unique to each character. This excerpt contains two characters: Lara, who is also the narrator of the story, and Ada, her landlady. Their ids, 'laraearly' and 'ada80s', are recorded in a separate

data frame (e.g., an Excel table), together with several character features, including name, gender, ethnicity and age. The latter is registered as either a numerical age or a life stage, as determined by the age scheme in Table 7.1, which is drawn up for the annotations and is based on theories by age critics Lorraine Green and Thomas Armstrong. Whenever the age of a character changes in the story, for example in a flashback or when the story spans several years, a new id is created to keep track of those changes. Where possible, the assignment of age relies on explicit mentions of the age or life stage of a character. According to Lara, Ada is "at least eighty" and her age is thus recorded as "eighties", a life stage defined as spanning from 80 to 89. If there is no explicit mention of a character's age in the text, age is derived from social markers and context. Later in Van Leeuwen's story, the reader learns that Lara had moved out of her parents' house six years earlier. Given that it is common in Western society for young adults to leave the parental home in late adolescence or early adulthood, and six years have passed, Lara was assigned the age 'early adulthood'.

Life stage	Corresponding numerical age
infant	0–2
child	3–11
● earlychild	3–5
● middlechild	6–8
● latechild	9–11
adolescent	12–19
adult	20–…
● earlyadult	20–39
● middleadult	40–59
● oldadult	60–79
● deepoldadult	80–…

Table 7.1: Age scheme including the life stages and their corresponding numerical age.

While context often provides the only indications of a character's age, there are downsides to depending on social markers. The designation of Lara's age relies on assumptions made about the age at which young people start living alone. The categories included in the age scheme are debatable and dependent upon social norms; it is also entirely possible that Lara only moved out of the parental home at a later age. The reliability of explicit mentions of a character's age by another character can also be questioned. Fictional characters, as well as the human annotators of the texts, might rely on their own understanding of age to assign it to other characters. For example, Lara, who is identified as probably being a younger adult, might be wrong when estimating the age of people who are more than twice her age. Similarly, narrators can be unreliable due to their lack of omniscience, personal involvement in the story, or because of a deliberate deception or questionable value-scheme, in terms of characterisation (Rimmon-Kenan, 103–4). These reservations have to be taken into account when relying on textual information to establish characters' ages.

To assign characters to the life stages of 'old adult' and 'deep old adult', the categories that are of interest to the analyses in this paper, the annotation relied on explicit markers of age, namely a numerical age of at least 60 and variants of the word 'old', and social markers, such as retirement and family relations, including parents of adults and grandparents. Some parts of this

method are again debatable. In real life, not all grandparents are older than 60 and so do not fit into the 'old adult' category. However, in children's literature, most grandparent figures are portrayed as being older than 60 (Joosen 'Grandparents'). This age norm in children's literature and in the annotation of the corpus for this article is self-perpetuating; as grandparents are mainly portrayed as older, all grandparent characters, except when explicitly assigned a younger age, are considered to be over the age of 60. Furthermore, relying on certain descriptions to assign the life stages of 'old adult' and 'deep old adult', such as the word 'old' or the physical feature of grey hair, can lead to circular reasoning when looking at the results; do older characters walk laboriously or are they labelled older because they walk laboriously? The interpretation of the results in the next section tries not to fall into this trap by taking the argumentations for the assignment of these categories into account.

After identifying all of the references to older adults in the 41 texts, all of the verbs and adjectives associated with, and words that stand in a grammatical possessive relation to (henceforth referred to simply as 'possessions'), these characters are extracted to the representation of these life stages on a large scale. To this effect, a Python script was developed in collaboration with TEXTUA, the Text Mining Centre of the University of Antwerp.[ii] The script uses the open-source software *spaCy* (Honnibal and Montani) to parse the syntactic dependencies in sentences. This results in a dependency tree, which links all of the words that are associated with each other, as well as the type of relationship they have; for example, a verb and its subject. By combining the resulting data from each text with the corresponding character list, the 'parser' can extract the relevant information of characters in different age groups.[iii] As the analyses in the next section seek to detect the differences between the representation of male and female older adult characters, group characters with mixed genders, e.g., 'pensioners', were removed from the data. After the extraction, all verbs, possessions and adjectives are lemmatised, a process which converts each word to its stem form, taking into account their part in speech. For example, depending on its intended meaning in the text, the term 'bears' would be interpreted as the plural form of the animal 'bear' or a conjugated form of the verb 'to bear'. After this process, the fifty most frequently used verbs, possessives and adjectives are extracted. While these lists are a great source of data for studying the representation of older people in children's literature, we need to take into account that a distorted picture may be presented. Extracting a single word removes possible negation and ignores nuance. To limit the bias that might arise from a naïve analysis of these words, the parser also extracts adverbs and adjectives pertaining to the verbs, possessions and adjectives associated with older age.

7.3 Results and discussion

While the complete collection of annotated texts fed into the parser contains 97 titles, data were collected from only 41 of those. Sixty percent of the titles did not include sufficient data on older adult characters; if older adult characters feature in those stories, their role is insignificant or they are mentioned only fleetingly as none of them undertake an action, are attributed with an adjective or have a possession. When looking at the distribution of older adult characters across books for readers of different ages, books for adults contain the most older characters; data were extracted from 11 of the 16 titles in this category. One possible explanation is that part of the intended readership of these books are older people and the other part are in the stage of life that precedes older age. Including characters who share certain characteristics with the intended readership, in this case their age, increases the chances of the reader's involvement with the

story. The same happens in children's literature, which usually features a large number of child characters. However, according to the most basic definitions of children's literature, the 'child' in the term refers to the intended reader. Thus, while researchers, including Barbara Wall and Perry Nodelman, argue that an adult almost always acts as a second reader or second addressee, these books are not initially written with (older) adults as the main target audience. Nonetheless, older adult characters do often play an important role in stories for children and traditionally have a close relationship with the child characters. In this corpus, however, the children's books contained the smallest proportion of older adult characters; only around one-third of the titles for seven and nine year olds contained relevant information about this life stage. The close connection between children and older adults is generally lost when the child ages into adolescence (Joosen 'Not Your', 184–85). While, from observation, one might derive that older characters will appear less frequently in books for young adults, just short of half of the books intended for young adults aged between twelve and fifteen feature older adult characters, a significantly larger percentage than the books for children.

In the next section, verbs, possessions and adjectives are discussed separately, according to the following structure: for each word type, several themes that emerged from the extracted word lists are analyzed, first in books for children, then in young adult novels and finally in adult novels. The themes that are included in the discussion are a selection of a handful of topics that were identified during a review of all six word lists for each word type, one for each gender over the three categories of books, according to the intended age of the reader. Some of these themes contain words that appear only in small numbers in the word lists but nonetheless contribute to the thematic cluster. The selection was made based on the relevance of a theme, as well as themes with the greatest number of words, with a preference for modified terms. A complete record of the word lists, including themes that are not discussed below, can be found in appendix E.

7.3.1 Verbs

Extracting verbs that are syntactically dependent on older adult characters from texts for children, young adults and adults provides an insight into the different activities in which this life stage is associated in fiction. Most verbs could be assigned to one of the following topics: the physical and mental modes of perceiving and processing surroundings, verbs that describe movement, ways of expressing oneself, leisure activities and living in older age with illness and mortality (see appendix E1 for a full word list). To keep the analysis of actions manageable and to the point, the following discussion covers all but the first of these themes. In the books for children aged between seven and nine, there is no large variety in the movements of older adult women; in terms of stationary actions, they 'stand', 'sit' and 'stop', while more dynamic verbs include 'come', 'pull', 'shuffle' and 'walk'. The latter is accompanied at least once by the adverbial phrase 'no longer well', which is very similar to one of the twenty most common words used by children aged between three and six to describe older people, as established by Allison Flamion et al.'s study, which included children aged between three and six and adults: "walks with difficulty" (Flamion et al., 7). This action, and more specifically its further description, suggests the stereotypical image of the slow older person walking with a laborious stride. Older male characters show a wider range of movement; they also 'stand' and 'sit', but their list of dynamic movements includes grander moves. Although they are not described as having more trouble walking, older men 'fall', which connects older age to frailty and pain. On the other hand, men

'lift (something or someone) high', which in turn indicates vitality in older age and contradicts the ageist image of the frail elder. Furthermore, older male characters 'sneak', 'disappear' and 'hide well', which may indicate a playful element. In children's literature, older adults engage in play often when accompanied by a child; the fondness of play shared by grandparent and grandchild figures are a common trope (Joosen *Adulthood*, 189).

In the entire corpus, women in the books for children are the only older adult characters whose speech is not modified; they 'say', 'talk', 'tell' and 'ask', but the manner in which they do so is not specified. The lack of information on how these characters express themselves is no doubt due to the small proportion of older female characters who play an active role in the stories for children. The only exception is 'strictly forbid', which evokes the image of a strict and stern older woman. This stereotypical character image is further constructed by one of the emotions older women in the books for children are said to display; they 'look angry'. However, this image is nuanced by the fact that they also 'look grateful' and 'become cheerful'. Modified verbs pertaining to verbal and non-verbal expressions of older adult male characters also construct different types of characters. One of the depictions outlined is that of the obnoxious, angry older man who 'says (something) gruffly and sternly', 'sighs deeply', 'laughs bitterly' and 'talks too loud'. Another character type emerges from the following modified verbs: 'whisper incoherently and agitated', 'talk very softly', 'say sadly' and 'look sad and dreamy'. These descriptions evoke the image of a frail and downcast older person. A more cheerful picture is painted by the last group of verbs: 'smile wide', 'put at ease' and 'look encouragingly and friendly'. A possible hint at the element of play, as discussed above, is also present in the word list of older adult male characters; they 'say conspiratorially'.

Although 'play with' shows up in the twenty most common words in children's descriptions of older adults (Flamion et al., 7), the references to play found in the actions of older adults in books for children remain largely implicit. Both genders do engage in 'drawing' and apparently do so 'well'. Other leisurely activities include 'baking' for women and 'cleaning' for men, subverting the gender stereotype of women being in charge of housekeeping. Several actions undertaken by older adult characters in the books for children refer to life, death and the deterioration of health. Both men and women 'die', but also 'exist' and 'live'. The latter is modified in the male word list by the adverbial phrase 'for so long already', which, together with 'be too old and too skinny', emphasises the characters' older age. In terms of health, older adult women 'breathe (ever more) slowly', while men 'feel weak'.

The discussion above considered only the children's books. In the books for young adults, the movement of older adult characters is presented in a slightly more stereotypical way than in the books for children. Apart from the static verbs, 'stand', 'stop' and 'sit', women 'trip backwards', 'shuffle softly' and it takes them a lot of effort to make certain moves; the verbs 'stoop', 'hoist' and 'bend' are all modified by the adverb 'laboriously'. Furthermore, there are traces of an element of anger in the way older people move, as women 'ram furiously' and men 'stamp frantically'. Many other actions related to the movement of male older adult characters emphasise the lack of it. Men 'lie dead still', 'sit motionless', 'lean' and 'stand encouragingly'. When they do move, they 'shuffle', 'walk', 'move' and 'turn around quickly', the latter being the only modified verb that hints at vitality in male older age.

A more diverse picture is painted of older adult characters in young adult novels when looking at the ways in which they express themselves and the emotions they display. There is a lot of overlap in the neutral ways older men and women say something; both 'ask', 'answer', 'tell' and 'say' something. The modifiers extracted for the latter, however, differ. Women are said to express thoughtfulness, urgency and kindness as they speak, while men are concise and despondent. Another difference found between the word lists of female and male older adult characters is the variety of verbs associated with both genders to express happiness and enjoyment, on one hand, and sadness and distress on the other. For the former, female characters only 'laugh' and 'giggle', while men 'smile' and 'chuckle' in addition to those. By contrast, as women 'hit', 'shout', 'wail' and 'cry', only the first two of these verbs are extracted for men. This suggests that older adult male characters in young adult novels are more often represented as being happy and older women are more often found in a state of distress, which is reminiscent of the stereotype of the hysterical hag, as identified by Susan Pickard. The adverbs accompanying the verb 'look' supports this hypothesis; women 'look anxious', while men 'look elated and serene'.

In addition to the references to more laborious movement, as discussed above, there are few references to illness in older age in books for young adults. In contrast to books for children, older people in these stories do not feel 'too old' or 'weak'. The only exception is that older women 'breathe excessively' and men 'leave (something) behind', which may indicate the making of a will due to a higher awareness of mortality traditionally attributed to people of advanced age. Leisurely activities, on the other hand, receive more attention in young adult novels, but are exclusively associated with male characters. They 'plant', 'build', 'drive' and 'knead', which are actions that fit in with the stereotypical activities that Patricia Crawford lists for grandfather figures in children's books, namely "taking walks or fishing" (Crawford, 162).

In the adult novels in the corpus, older adult characters are represented as engaging more often in communicating and expressing their emotions than in active movement. When they are active, older characters are shown interacting with objects more often than in the books for young readers. This can be inferred from the high number of verbs of object movement in the fifty most common for both men and women. More so for older men than for older female characters, adverbs emphasise laborious movements. Men 'push', 'pull close', 'slide far and open', 'throw hard' and 'stroke rhythmically', but they also 'shift' and 'drag', both 'laboriously'. In the list of verbs associated with female characters, there is no indication that they have difficulty moving objects; they 'pull', 'push open wider', 'shake vigorously' and 'close well'. As in the books for younger readers, there are also verbs that refer to static and dynamic postures or movements of older adults, including 'stand', 'sit' and 'lie', as well as 'walk' and 'step', respectively. However, there is less diversity in these types of actions in the adult novels. A second difference is that, with the exception of 'walk slowly' for older men and 'step carefully' for women, there is no reference to a decrease in mobility.

Whereas the books for young readers relate more verbs referring to movement with older age, a large proportion of the fifty most common verbs in the adult titles refer to expression and emotion. The most variety is found in the word list of older female characters, with both verbal and non-verbal expressions, as well as a large range of emotions (see appendix E1 for full list). To select a few, women 'say' things 'softly', 'angrily' and 'cheerfully', 'speak very articulately' and 'greet briefly'. A number of other verbs hint at a negative image of older age; women 'call

desperately', 'shout anxiously' and 'helplessly', 'feel lonely', 'miserable' and 'restless' and 'become uncomfortable'. Furthermore, some adverbs evoke the ageist image of the rambling granny: 'tell imperturbably', 'long-windedly' and 'talk incessantly'. Male older characters are depicted as less communicative in books for adults and, together with the emotions they display, the image of an angry older man appears. Older men 'say directly' and 'conspiratorially', 'interfere', 'shout', 'cry', 'become mad' and 'grumpy' and 'look suspicious'.

Adults, more so than children, associate older age with "disease", "health", "dependence" and "frail[ty]" (Flamion et al., 7). A parallel observation can be made for the representation of matters of life, death and health in the adult novels in the corpus when compared to the books for younger readers. While in the latter, references to illness are made rather subtly (their breath comes 'ever more slowly' or 'excessively'), the former explicitly describes older adult characters as 'feeling sick' and 'looking sick', both of which are modified verbs found in the female word list. Furthermore, the advanced age of both genders is suddenly seen as unduly high as they 'become too old', 'live too long' and 'feel ancient', as opposed to just 'long' or 'old'. It is perhaps not surprising then that these 'too old people', who have already been shown not to have a wide range of motion in books for adults, are not often depicted performing a leisurely activity. Moreover, the few activities older women engage in paint a stereotypical picture of their occupations; they 'polish shinily' and 'scrub too hard'.

7.3.2 Possessions

The possessions of older adult characters that the parser extracted from the corpus mainly relate to the following topics: body parts, health, personal relationships and community, home and household objects and clothing. The discussion in this section will focus on the first three topics (see appendix E2 for wordlists of all five topics). In the books for readers aged between seven and nine, for all of these themes, there is more variety in the possessions of male characters and they are modified more often. The only parts that are described are 'head' and 'hair', which are, respectively, 'small' and 'long'. The latter stands in opposition to the stereotypical image of older women with short hairstyles, being short hair or a bun, found by Villar and Fabà to be prevalent among primary school children. What is strikingly missing as a modifier of 'hair' is its color; 'white' or 'grey'. This familiar image was found by Flamion et al. as one of the twenty most common words children use when describing older people, but only appears in this corpus in the adult novels. It is possible that this information is omitted from the text of children's books because it is clearly visible in the illustrations included in or on the books. More so than for females, male body parts extracted from the books for children point toward an emphasis on the deteriorating body in older age. Older men are described as having a 'skinny body', 'skinny hands', 'old mouth' and 'old tongue', 'bony fingers', 'unwashed skin' and 'wrinkles'. The same theme of deterioration, again mainly present in male possessions, is found in words associated with health and attributes used to improve quality of life. While 'life' is included in possessions of both genders ('whole life' in the male list), the more abstract 'existence' and the reference to mortality of 'grave' feature only in those of older male characters. Both older women and men seem to wear 'glasses', but only the latter require 'dentures' to live their lives comfortably. The encompassing word 'health' also appears only as a possession of men, indicating that this is not only a theme for those characters, but is also thematised in the children's novels.

Looking at grammatically possessive relationships, family members account for the largest proportion of people in the communities of older characters in the books for children. This is mainly the case among men, while the relationships of older women appear to be more varied. In the list of women's possessions, the following people outside the family sphere are named: neighbor, locksmith, student and class. The latter two suggest an educational work environment, contradicting the image of the retired older person who disengages from professional communities. This stereotypical image fits within disengagement theory, constructed in 1961 as part of social gerontology, which states that it is the norm for older people to withdraw from social relationships and suffer a decrease in societal roles (Cumming and Henry). Although this theory was soon criticised, partly because it encourages ageism and partly because the statements are not supported by empirical research, this negative view of older adults often creeps into the way this life stage is considered in contemporary society. In 1990, Fraboni, Saltstone and Hughes developed the Fraboni Scale of Ageism, a method to measure this attitude towards older people. One of the items on which their scale is based is the following: "Many old people are not interested in making new friends, preferring instead the circle of friends they have had for years" (Fraboni et al., 62). Margaret Morganroth Gullette argues that generalisations such as this are largely accepted by young people (Gullette *Ending Ageism*, 58). The books for children aged between seven and nine studied in this article seem unaffected by this aspect of ageism, as in the list of possessions associated with older male characters, the only person mentioned who is not a relative is a 'new friend', breaking the image of the disengaging older adult. This observation was also made by Joosen ('Second Childhoods'), who analyzed four children's books from the Netherlands and the UK in search of the representation of older people according to four ageist tropes.

In parallel to the books for children, the young adult novels included in the corpus (intended for twelve-to-fifteen year olds) include more data on older male than female characters, albeit with a smaller distinction. Most of the body parts named of older characters do not immediately evoke stereotypical images. With the exception perhaps of 'lap' and 'joint' (both from the male list), which can be argued to evoke a predominantly older image, words such as 'shoulder' and 'rib' can be associated with any age. However, when we look at the adjectives used to describe body parts, patterns emerge for both genders. On the one hand, older men have 'thin hair', a 'wiry body' and 'bony hands'. The latter is also included in the female list, together with 'dry hands' and 'last breath'. On the other hand, older women in young adult novels can also have 'fleshy arms' and 'chubby fingers', suggesting possible obesity. Although this interpretation makes it seem as though older people are depicted as being either under- or over-weight, this contrast is most likely created by the nature of this study; by mainly focusing the discussion on possessions attributed with an adjective, references such as 'arms', 'legs', etc., without an adjective, which are also present in the data, fade into the background.

No adjectives are included for the possessions pertaining the health of older adult characters in the books for young adults. Based on gender, two distinct images of older people occur; the female list contains the words 'death' and 'deathbed', while the male list features the more positive words 'movement' and 'wheelchair'. The latter emphasises the attributes that older people have at their disposal to stay mobile. In Flamion et al.'s study, three of the twenty most frequently used words to describe older people pertained to assistive devices. Adults barely, if at all, referred to such attributes. In this regard, the construction of older age in the books for young adults in this corpus is more closely related to children's perceptions compared to those of adults.

In terms of the community around older adult characters, there is relatively less variety in the young adult books than in the titles for children discussed above. The list of female relationships is short ('man', 'daughter' and 'friend'), while that of male characters is slightly more varied. In addition to predominantly family members, there are also references to older men's 'former people' and 'first customer'. The latter is reminiscent of the 'new friend' in the books for seven-to-nine year olds that indicates a wider social circle than just family.

In the adult novels included in the corpus, there is less difference in the proportion of data from men and from women, and possessions are described more often than in the books for younger readers. The characterisation that emerges from these words is predominantly stereotypical. Although few body parts are attributed with an adjective in both the male and female lists, older age is similarly represented in both genders. Older women have 'dry eyes', a 'worn-out' and 'rigid body', an 'uncooperative' and 'stiff back', 'old fingers' and 'grey', 'tangled', 'thin white' and 'uncombed hair'. Men are described as having 'blotchy hands', 'cold cheeks', 'frail' and 'weak legs', a 'hairy' and 'hunched back', 'yellowish whites of the eyes' and a 'heavy', 'unwieldy' and 'slow body'. This emphasis on the deteriorating older body is reflected in the adult descriptions of older people in Flamion et al.'s aforementioned study, where 'frail', 'disease', 'death' and 'health' can be found among the twenty most common words. On the other hand, however, positive words such as 'experience', 'wisdom', 'kindness' and 'love' can also be found in their descriptions, which do not feature in the lists of possessions extracted from the adult books in the corpus. The words associated with health and the associated devices paint an equally bleak picture; older female characters wear 'thick glasses', but beyond that only references to 'death' and 'deathbed' are made. The former also features in the male list, together with 'birthday'.

From what we can derive from the possessives, the community around older people in books for adults is significantly limited and almost exclusively includes close family members. The only exception is found in the male list, which includes 'murderer', linking older age to mortality. Furthermore, the only mention to a dead relative occurs here, as one older adult male character is described as having a 'dead wife'. Novels for adults seem to be more explicit in their connection of ageing and death, as "children's books are quite selective in the topics they address" (Joosen *Adulthood*, 24). A further gap that can be identified in the texts for young readers in this corpus and is filled in the adult novels is embodied in the 'absent daughter' as a possession of older female characters. In her discussion of senior parents and their adult children, Ellyn Lem states that caring for an ageing parent can be a complex responsibility and carries the risk of leading to estrangement (Lem, 30). The resulting loneliness suffered by the parent "is often not communicated due to a lingering stigma attached to it" (ibid., 42). While absent adult figures, often parental, are a common trope in children's books, for the benefit of the child-grandparent relationship, their absence is not usually labelled as such. The systematic equation of older adulthood and loneliness and the deterioration of their social relations, as discussed above, is confirmed by Flamion et al., where the most frequently occurring word with which adults describe older people is 'loneliness' (Flamion et al., 7). Standing close to older adulthood as their next phase in life, being abandoned by their own children seems to be a fear that adult authors are more likely to express in books for adults than for children.

7.3.3 Adjectives

Although the analysis of adjectives probably comes to mind first when wanting to study characterisation in fiction, this category contains the least amount of data. In her monograph on characterisation in children's books, Nikolajeva identifies gender stereotyping in the direct description of characters, where mainly female characters are described, as opposed to men (*Rhetoric*, 189). However, this is not reflected in the corpus of this article. Mainly in the children's books and books for young adults, there is more variation in the adjectives associated with men. This may be due to a possible smaller number of female older adult characters that play a role in those books. Across the genders, most of the adjectives associated with older adults pertain to one of the following categories: appearance, emotions, character traits, illness and age (see appendix E3 for the full word list). In the discussion below, the first three will be examined in more detail. The latter is not included in the analysis, as both genders in all three groups of texts (children's literature, young adult novels and adult novels) are described mainly as 'old'. It is very likely that this description is an identifier rather than a description; during the annotation of the texts, age is assigned to characters based on explicit age markers, of which 'old' is the most obvious one for older adulthood. Characters are thus categorised as older people by the presence of the adjective 'old', rather than the adjective being a feature of characters otherwise considered to be older adults.

In the books for children aged seven to nine, female older adult characters are stereotypically described only as 'grey', while men are 'dusty', 'small' and both 'fat' and 'skinny'. Apart from looking angry and grateful, which was discussed in the section on the actions of older women, no emotions are attributed to these characters in the books for children. Older men display a wider range of emotions, both positive and negative; they are 'not scared' and 'maybe happy', as well as 'sad'. Furthermore, men are often depicted as being bitter characters in the children's books in the corpus. The adjective 'angry' is included in the word list several times and is further described with the modifiers 'maybe', 'very' and 'also'. In terms of character traits, this characterisation emerges only from the adjective 'strict'. Other direct descriptions of older men are much more positive: 'kind', 'well-behaved', 'clever', 'valuable'. It is also this picture of older adulthood that is painted from the description of children, as observed by Flamion et al. References to negative emotions do not appear in the twenty most frequent words of the participants in that research, but 'kind' does (ibid., 7). Again, there is less data for older female characters; their character is described as 'patient', 'perhaps more interesting' and 'thrifty'.

In the books for young adults, the emphasis in the direct descriptions of older adult characters lies mainly on their appearance and, for women, on their personality. From the former, the stereotypical image of the deteriorating body of the aged person emerges for both genders. Women are 'small', 'grey' and 'wrinkly', while men are 'skinny', 'white' and 'thin'. One male character is also described as 'unwashed', which may indicate bad care. The same character is described with the only emotion associated with older male characters in the books for young adults, 'dissatisfied', which stands in contrast to the emotions of serenity and elation assigned to these characters in the analysis of verbs. Similarly, for older women in the young adult novels, only one adjective referring to an emotion was extracted from the corpus, which also goes against their display of anxiety previously discussed; they are depicted more positively, as being 'happy'. For women, more data were extracted pertaining to character traits, but they are

represented in a similar vein, as being ‘tough’, ‘always kind’ and ‘honest’. The single character trait mentioned as an adjective for older adult men in the young adult books is ‘secretive’.

From the adult novels, roughly the same number of total adjectives were extracted for both genders. In terms of appearance and emotions, the description of female older adult characters is more diverse, while men are attributed with a wider array of character traits. None of the adjectives describing the appearance of older people refer to the stereotypical image of the older body as being in a state of deterioration. Women are ‘flat’, ‘already a bit less blotchy’, ‘tall’ and ‘sweaty’. Men are also described as ‘tall’, as well as ‘heavy’. In terms of emotions, the list of extracted words associated with older women contains contrasting adjectives, painting a diverse picture of these characters. They are both ‘always happy’ and ‘not happy’, ‘very upset’ and ‘very relieved’, ‘grateful’ and ‘desperate’. The list of men’s emotions is shorter, containing only ‘happy’, ‘angry’ and ‘proud’. Finally, the character traits of older adult characters in the adult literature reflect an ageist trope for each gender. In her study of the history and meaning of fairy tales, Marina Warner identifies the prying older woman as a stereotypical character that has long since existed in literature (27–50). Older female characters being described solely as ‘very curious’ in the adult novels might allude to this image. Furthermore, in addition to being ‘interesting’, ‘good’ and ‘not climate neutral’, older men are associated with the character trait of wisdom. Taken together with the lack of direct descriptions of older men referring to declining health, both physical and mental, this association seems to exemplify the ageist archetype of the wise older man as a denial of the decline narrative. This carries forth “the idea that with old age comes wisdom and that the old still have their knowledge and experience to offer to younger generations” (Joosen *Adulthood*, 194). Flamion et al.’s study confirms that this stereotypical equation of older age and wisdom is also made by adults in real-life situations. Among the top twenty words used by adults to describe older people are both ‘wisdom’ and ‘knowledge’ (Flamion et al., 7).

7.4 Conclusion

Several conclusions can be drawn from taking the observations on verbs, possessions and adjectives together. With regard to gender differences, older women’s mobility in the corpus of Dutch texts is underrepresented when compared to that of older male characters, but the laborious movements described for the latter category both confirm and contradict physical deterioration. Words referring to illness are associated more often with the male figures, but they are generally also attributed with a larger variety of assistive devices, suggesting that the effects of a deteriorating body can, in part, be overcome. Contrastingly, the range of expressions, character traits and emotions attributed to older men is less diverse than that of women in this corpus, portraying them as flatter and less communicative characters. A predominant emotion among men seems to be anger; the stereotypical image of the angry older man surfaced several times in the analyses. Other stereotypes identified in the corpus are the wise older man, the nosy older lady and the distressed, rambling granny. In contrast to this image of the unpleasant older man, many more elements that can be associated with play emerge for this gender, facilitating the connection between male older age and childhood.

Looking at differences between texts intended for readers of different ages also reveals certain patterns, but maybe more significantly, several gaps. While age is considered as a significant characteristic of older adults equally in the books for children, young adults and adults, the

treatment of this characteristic differs. In the former, an emphasis is placed on the length of the lives of older adults, but no association with mortality emerges from the verbs, adjectives and possessions associated with older adults in the corpus. This shifts in the books for adolescents, where references to death are more common, for example in the verb 'leave behind'. In the adult novels in the corpus, the age of older people is sometimes seen as unduly high, as they 'become too old' and 'live too long', solidifying the connection between ageing and dying. The adult novels also portray older adulthood as a stage in life where people are prone to illness and experience the effects of a deteriorating body. While this is a realistic, if perhaps exaggerated, re-presentation, older people in the books for children are less often depicted that way, pointing to a gap in the books for younger readers in the corpus. The image of the deteriorating older body is, however, included in the young adult novels. Moreover, the analysis suggests that older characters in the adult novels have few connections with people outside their household and almost none with people outside their close familial circle, constructing a sense of loneliness.

While the dense summary given above attests to the benefits of applying digital tools to the study of the construction of older age in literature for readers of different ages, the method did not pick up on the progressive themes that Dutch children's literature is known for—this can be either because the books do not display this trend or because the themes are expressed in ways that the method does not include (e.g., illustrations, metaphors, more subtle suggestion). Based on the occurrence of verbs, possessions and adjectives, it seems that illness and death featured less prominently in the books for children than those for older readers, as did loneliness in older age. Finally, the analysis did not take into account the elements of direct speech and focalisation. All of the descriptions of older characters were treated equally and not called into question, although, in a story, it can make a difference whether an ageist remark is made by a likeable character or by an antagonist. Although the texts in the corpus have been annotated not only to identify references to characters but also to attribute speech instances to characters, the scope of this article did not allow for including this extra layer in the analyses. However, an analysis focusing on only one category of texts, for example adult novels, could incorporate this element.

In a discussion of the significance of positive images of ageing in early children's literature, McGuire cites that it is important for educators to first "evaluate one's own attitudes about aging. Ageist attitudes are contagious and children will assimilate the attitudes of those around them" (2605). I argue, with this article, that this susceptibility to a negative attitude towards older age does not stop when childhood is over. The promotion of positive aging, as important as it is for children and their interaction with older people in their daily lives, is equally important for adolescents and adults at different stages. On one hand, adolescents, in their search for identity, are more so than children confronted with the fact that soon they will be part of 'the adult world' in all its facets and that they, too, will one day become older. On the other hand, most adults will, at a certain point, suddenly have to carry the responsibility of caring for their own parents, as well as thinking increasingly about their own older age. The consequences of negative perceptions of older age among adults are abundant. To name just two, adults might develop a fear or unwillingness to age, as well as exclude older people from situations that are about them. For example, Joosen noted that, in 2020, it took months for an older adult to appear in a Flemish current affairs program about the recent corona pandemic, when they were the ones hit hardest (Joosen *Hoe oud*, 6). It is also noteworthy that, while panels in such programs and in political debates, which are often organised by younger adults, increasingly pay attention to a proper

representation of men and women, supplemented by spokespeople for minority groups, they further fail to take into account an equal distribution of the age of those participants.

Large-scale studies into the representation of older adulthood in fiction have an important role to play in shifting Western thinking towards a more diverse representation of older age, which is an important step towards active aging, a paradigm that recently emerged in gerontology that considers aging from a positive perspective (Fernández-Ballesteros et al.). It is, however, important that this movement does not only take diversity on an individual level into account, but it also should not neglect the advances with regards to gender inclusion. The analyses documented in this article are the first to use tools and methods from the field of digital humanities to investigate the representation of female and male older adult characters in children's literature, young adult fiction and adult novels in a comparative study. Due to the combination of a fairly large number of texts, namely 41, studied simultaneously through a fine-grained approach of digitally extracting words associated to characters in a specific life stage, patterns and gaps could be identified in the comparison based on gender, as well as in the different categories of texts according to the age of the intended reader.

The fact that ageist representations of older adults seem to appear predominantly in the adult books in the corpus calls into question why, in the effort to eradicate such representations, there are not more studies into the traces of ageism in these books. While the arguments that children assimilate ageism at an early age (Montepare and Zebrowitz) are valid and the call for research in children's literature remains important, I think there is still much to learn from studying books for young adults and adults. To carry out further research on a large scale, the methodology of this article is a useful tool for studying certain facets of characterisation, especially if it is matched with analysis through close reading, which helps to contextualise the results for specific titles. By enabling the extraction of a lot of data, the method can help researchers to identify trends and select themes to explore further, taking us closer to diversifying the images of older adults in literature.

[i] Fragment taken from *Alles Nieuw* by Joke van Leeuwen. Page numbers are not included in this article because most texts were obtained in eBook format. All translations from Dutch are made by the author. Original text: "Dit huis is honderd jaar, zegt de huisbazin, die volgens mij minstens tachtig is. Allemaal van voor mijn tijd. Dit is mijn tijd. Ze woont beneden, ik moet haar elke maand contant betalen, dat noteert ze in een schrift, turend naar het puntje van haar pen, terwijl ik naar haar handen zit te kijken, naar de landkaart van aders en donkere vlekken die rivieren en dorpen kunnen voorstellen."

[ii] The code is available in open access via the following link: http://textua.uantwerpen.be/app/cafyr. A prototype of this parser was used by Vanessa Joosen in 'Research in Action: Constructing Age For Young Readers' (2021) to study adjectives in Philip Pullman's *La Belle Sauvage*. See this article for a brief introduction of previous studies conducting analysis with other implementations of automatic syntactic dependency parsers.

[iii] These, as well as all following words extracted from the corpus are translated from Dutch by the author. Any mistranslations pertaining to meaning or underlying tone of the terms are their own.

Chapter 8 General Conclusion

8.1 Main findings

Reading the Reader from a Distance has examined how the construction of age in fiction is shaped by the age of the intended reader, delving into the intricate dynamics of age on two levels: the surface of the text and the semantic layer in the text. By focusing on the interaction between surface features and the age of the intended reader, as well as fictional characterisation and metareflections about age, this dissertation has provided a comprehensive analysis of 738 texts in two languages, intended for readers with a minimal age range spanning from 7 to 18 years old, and encompassing a total of 25.307.024 words, which amounts to approximately 23 times the total word count of the *Harry Potter* books. The aims of this research were twofold: firstly, to contribute to a closer understanding of the construction of age in the corpus in terms of formal (stylistic) features and literary representations of different age groups, and secondly, to compile a set of methods for digital text analysis that enables the study of these aspects. The central argument put forth in this dissertation is that both explicit and implicit ideologies related to age are present in books intended for readers of all ages, but the way in which these ideologies manifest themselves varies depending on the intended reader's age. While contributing to the ongoing academic tradition of critiques of literary representation and the comparison of texts for children and adults, *Reading the Reader* extends beyond this dichotomy by recognising the segmentation inherent within children's literature, which is primarily based on the age of the intended reader rather than factors such as gender or ethnicity. The findings of the publications in this dissertation are synthesised in the following section to provide a comprehensive overview of the research and a general conclusion for each of the research objectives and questions outlined at its outset. To close off this section, I will reflect on the implications, strengths, and limitations of the research conducted in this dissertation, based on which I will offer recommendations for further research.

8.1.1 RO 1: gaining insight into the construction of age

The first research objective of this dissertation was to gain deeper understanding of the construction of age in literature for young readers. This objective was approached through two research questions, each addressing distinct aspects of age construction. Research question 1.1 investigates the impact of the age of the intended reader on the construction of age *by* a work of fiction. This question primarily focuses on the characteristics of texts, encompassing formal features, textual complexity, and trending topics. The analyses contributing to this part of the research sought to reveal some of the expectations that adult producers, including authors, editors, publishers, as well as mediators such as bookstores and libraries, hold for readers of specific ages, as reflected in the text.

Chapter two examined the oeuvres of ten Dutch and English crosswriters in a case-based manner; employing methods from the field of stylometry, the oeuvres of several authors (Guus Kuijer, Hilde Vandermeeren, and J.K. Rowling) were considered individually and all English titles were analysed simultaneously. Chapter six focused exclusively on the oeuvre of David Almond,

employing stylometry as part of its analyses. Both sets of analyses demonstrate that the age of the intended reader has an important influence on the stylistic clustering of the works of the crosswriters included in the corpus, and chapter two suggests that stylistic similarities are generally more correlated with the age of the intended reader than with the author's identity or year of publication. However, the publication date helps to explain outliers in age-related clustering. Furthermore, the fine-grained analysis in chapter two reveals that texts intended for a young adult readership show greater stylistic similarity to adult works than to books for older children. The case study considering Hilde Vandermeeren's oeuvre revealed a distinct progression in her writing style, transitioning from texts intended for young children to older children and ultimately to adult readers. However, it is important to note that such a clear pattern does not appear uniformly across all cases, as evidenced by the results of the stylometric analyses in chapter six.

Chapters three, four, and five analysed additional elements of writing style beyond those investigated with stylometry, including average word length, sentence length, lexical diversity, and ratio of dialogue to narration. Chapter three focused on J.K. Rowling's *Harry Potter* series, examining the correlation between the series' increasing complexity in form and content and the age of the intended reader, as established by different institutions and readability formulas. The series displayed a notable increase in difficulty based on formal aspects like sentence length and subordinate clauses, while lexical diversity showed a more modest rise throughout the series and did not reach the value of this measurement of Rowling's adult novels. The analysis of evolving topics revealed a shift from childhood-oriented themes to more mature themes, with a noticeable increase in the theme of 'death' suggesting a change in implied readership. Chapter five analysed the works of Joke van Leeuwen, revealing formal distinctions between her books for different age groups, particularly in terms of word and sentence length as well as lexical diversity, which all tend to increase with the age of the intended reader. The ratio between indirect speech and dialogue also distinguished her work for children from that for adults. In chapter four, a statistical analysis was conducted on the entire corpus of Dutch and English texts, indicating that the ratio of direct speech to narration was not significantly correlated with the age of the intended reader in either language. However, significant positive correlations were found between the age of the intended reader on the one hand and average word length, average sentence length, and lexical diversity within the corpus on the other; as the age of the intended reader increases, so do the values of these formal features.

The second research question contributing to the first aim of this dissertation explored how the age of the intended reader influences the construction of age *in* a work of fiction, particularly regarding the representation of characters through their speech, actions, direct descriptions, and possessions. By examining the traits assigned to characters of different ages, these analyses contributed to insights into broader cultural and social beliefs surrounding age. The analyses followed a bottom-up approach, starting from the information contained in the text and employing a structuralist and semiotic paradigm to identify recurring patterns that operate within and between the texts, considering the way in which age norms are established through language associated with representations of fictional age. Annotated material was used to collect relevant data from a subset of the corpus.

The ideological aspects implicitly expressed in the oeuvre of Joke van Leeuwen were studied in chapter five through the presence of characters from four different age groups (child, adolescent, adult, and older adult) and the proportion of their speech in the total number of words in direct

speech. As the age of the intended reader increases, there is a clear decrease in the proportion of child characters and their speech while the share of adult speech increases. Children, who have a louder voice in Van Leeuwen's children's books, speak significantly less in her books for adults. The representation of adolescent, adult, and older adult characters remains stable throughout all categories, but the share of their speech grows as the age of the intended reader increases. The expression of explicit age ideologies through the metareflections extracted from this subcorpus shows less variation based on the age of the intended reader, especially in the portrayal of older adulthood.

A large part of chapter six was devoted to the construction of fictional age in the oeuvre of David Almond. In terms of the speech of fictional characters, Almond's books intended for readers of different ages showcase variations while maintaining recurring themes. Colloquial language is prevalent across all books, but the specific words used differ. His adult, young adult, and children's books feature young and adult characters discussing darker themes, but from what we can derive from the most distinct words in their direct speech, violence is more graphically depicted in Almond's adult and young adult novels, and less so in books for children. The theme of adult responsibilities, work, and play emerges in both Almond's adult and young adult fiction, with distinctions between child, adolescent, and adult characters. In his adult novel, child and adolescent characters emphasise parental and student/teacher relationships, while in his young adult novels, adolescent characters explore conflict with parental authority and interactions with peers. Overall, Almond's books for readers of different ages feature recurring themes such as community, love, and cruelty, but also exhibit variations in vocabulary and topic prominence. This chapter also considered characterisation through verbs, adjectives, and possessions in a grammatical sense. Characters of a specific age were regularly described similarly in books intended for readers of different ages. For example, the image of the rebellious and socially alienated adolescent occurs in both his young adult and adult books. However, there was also evidence of more diverse depictions, for example that of childhood; in his children's books, the physical size of child characters as it is presented in the most frequent adjectives extracted from the subcorpus is negatively connotated with words such as 'ickle' and 'scrawny', potentially reflecting how children talk about each other, while in Almond's young adult novels, this characteristic gets a more endearing undertone by the use of 'delicate' and 'precious'.

Chapter seven focussed on the representation of older adulthood in 41 Dutch books intended for readers of different ages. Here, the examination of the representation of older adults highlights differences in the treatment of advanced age across books for children, young adults, and adults. While the emphasis in the children's books lies on the longevity of older adults' lives without explicit connections to mortality, references to death become more prevalent in adolescent literature, particularly in verbs such as 'leave behind'. Adult novels in the subcorpus of this chapter often depict advanced age as excessively burdensome, emphasising individuals becoming 'too old' and 'living too long', thus solidifying the association between aging, decline, and dying that is missing from the books for children in this section of the corpus.

8.1.2 RO 2: compiling a set of digital text analysis tools

The second research objective of this dissertation was to compile a non-exhaustive set of methods for digital text analysis that enables the study of age-related questions in fiction for young readers. The compilation of such a set of methods has several important benefits. Firstly,

they address a significant gap in existing children's literature research, which has been limited in scale and scope. Second, while digital analyses have been applied to children's literature in a handful of studies, the application of more than one type of analyses, let alone of a comprehensive set of methods, to a single corpus is still rare. By bringing together various approaches, this research contributes to a more holistic understanding of age-related themes in children's literature, young adult literature, and fiction for adults. Additionally, this dissertation aligns with the increasing use of digital text analysis in children's literature studies, expanding the methodological toolkit and fostering a more nuanced understanding of age-related questions in the field. This objective was supported by six research questions. The first five research questions aim to investigate whether the application of various digital tools for text analysis can uncover distinctions between texts intended for readers of different ages, with a focus on explicit and implicit age ideologies. The tools used in this analysis were chosen based on their potential to uncover implicit and explicit age ideologies in fiction, both in terms of the intended reader and the fictional age depicted. They are drawn from subfields of computational literary studies, including stylometry, readability analysis, topic modelling, and syntactic parsing and cater both to quantitative analyses of large corpora and the study of patterns in smaller datasets. By incorporating these tools, the research expanded beyond the traditional methods of literary analysis and sought to establish an approach that considers close reading and distant reading as complementary. As is apparent from the discussion in the previous section, the application of these methods to the corpus result in interesting insights into age ideologies, but there are certain limitations to the methodology, which will be elaborated on in the following section. The last research question takes into account the poetics of the authors in the corpus when studying their works and explores whether a digital analysis comparing works of crosswriters reflects their views on writing for readers of different ages or reveals other patterns.

To examine characteristics of the texts in the corpus, stylometric analyses were used in chapters two and six, as well as analyses of surface features in chapters three, four, and five. The conclusions drawn from chapters two and six suggest that stylometric analyses can indeed reveal formal distinctions between texts intended for readers of different ages (RQ 2.1). The analyses in chapter two also consider the identity of the author and the time of publication, but almost all stylistic clustering can be explained by the age of the intended reader. By extracting a list of most frequent words and manually culling it to remove pronouns, both chapters demonstrated that stylistic similarities can be found in the frequency distributions of function words of the books in this corpus intended for readers in the same age categories. The analyses of surface features conducted in chapters three, four, and five laid bare several distinctions between texts intended for readers of different ages (RQ 2.2). Although assigning age markers to the individual books in the *Harry Potter* series is problematic due to their crossover nature, digital analyses were able to detect age-dependent characteristics to some extent. Using statistical linear mixed models to analyse the entire corpus of this dissertation, chapter four argues that, while token-based lexical features, such as average sentence length and word length, have been questioned as reliable features for authorship attribution (Yule, and Tallentire *Appraisal*), they are more closely linked to categories based on the age of the intended reader. Additionally, lexical diversity showed a strong correlation with the age categories. The ratio between dialogue and narration does not exhibit a significant correlation. The findings from chapters two through six contribute to a deeper understanding of the interaction between writing style and the intended reader, providing reliable features for readership attribution. Additionally, based on the limited use of topic modelling in this dissertation, a tentative but noteworthy conclusion emerges, suggesting that

this approach holds potential to lay bare thematic distinctions between texts intended for readers of different ages (RQ 2.3).

Turning to the study of the construction of fictional age in the text, in chapters five, six, and seven I have examined annotated texts with a specific emphasis on age using various digital tools. The research in chapter five showed that the examination of character distributions and the proportion of their speech, on the one hand, and metareflections on the other, can reveal differences in implicit and explicit ideologies, respectively, in the oeuvre of Joke van Leeuwen (RQ 2.4). The analysis of implicit age ideologies involved the enumeration of characters included in the character lists and the extraction of their direct speech from the accompanying annotated texts. Additionally, metareflections along with information pertaining to the age of the character making the statement, were interpreted through a close reading approach. The use of a syntactic parser that utilises part-of-speech tags and grammatical dependencies in chapters six and seven proved valuable in gaining a greater understanding of the influence the age of the intended reader has on the construction of implicit age ideologies in the oeuvre of David Almond and in 41 Dutch-language books, respectively (RQ 2.5). By investigating verbs, adjectives, and possessions (in the grammatical sense) associated with fictional characters, patterns emerged in the representation of different age groups across a large number of texts. Verbs predominantly fall into the categories of verbal and non-verbal expression, movement, leisure activities, sensory perception, and processing of the environment. Adjectives primarily describe physical attributes, character traits, emotions, illness, and age. Possessions are mainly related to the human body, work, health, personal relationships and community, home and household objects, as well as clothing. The patterns identified in these subcorpora enable meaningful comparisons across texts intended for different readers, revealing nuanced variations in the portrayal of specific categories. For example, the deteriorating body of older adults is discussed in books for all young readers in the subcorpus of Dutch annotated texts, but these characters are depicted having health aids more often in the children's books than in the books for young adults.

Certain chapters in this dissertation expanded beyond an exclusively textual analysis framework to incorporate the authors' poetics regarding writing for readers of different ages (RQ 2.6). While studies on the formal aspects of adult literature often assume that an author's writing style is unique and manifests itself subconsciously, some authors of children's literature may consciously adapt their writing for younger readers. However, other authors deny consciously adapting the content or writing style of their works based on the intended readership or claim not to write with a specific age in mind. The poetics of author Hilde Vandermeeren, which she says are based on her experience as a teacher and mother as well as on her degree in psychology, align with the results of a stylometric analysis of her works conducted in chapter two. Vandermeeren's emphasis on understanding the perspective of children, her consideration of the age of the intended reader during the writing process, and the involvement of her own children in providing feedback on her drafts corroborate her stated approach to children's books. The analysis confirms that she adapts her writing style for readers of different ages, revealing a distinct division between books for young and older children, as well as adult readers. Additionally, the analysis indicates that her books for children exhibit subtle stylistic features that correspond to age, surpassing conventional distinctions based on genre or publication date.

Also discussed in chapter two were the poetics of Guus Kuijer and Philip Pullman, both of whom reject strict boundaries between children's and adult literature. The stylometric analyses of their

works, however, paint a different picture, which may indicate a more active role of their publishers in determining the age of the intended reader of their books. Kuijer's books show subtle stylistic tendencies that correspond to different intended readerships, albeit with certain outliers. An additional analysis highlights the deliberate stylistic choices made by Kuijer that can be associated with the intended reader, such as narrative tense and the (lack of the) physical nature of actions, extending beyond function words to content words. Pullman asserts that he writes solely for himself and that if his stories appeal to children, that it is a welcome outcome but not a deliberate aim ('Questions'). However, he does acknowledge the role of publishers in categorising his works, such as the *His Dark Materials* trilogy, which was marketed for young adults. In the stylometric analysis, we observe an evolution in Pullman's writing style, with a growing distinction between his texts for young adults and adults compared to those for young children. While Pullman may not consciously write for a specific age group, a discernible stylistic variation for younger readers emerges throughout his career.

Chapters five and six presented digital text analyses on the oeuvres two authors in the corpus; a study of surface features in the works of Dutch author Joke van Leeuwen and a stylometric analysis of British author David Almond, respectively, and included a brief discussion of their poetics. Van Leeuwen has expressed a reluctance to assign specific age labels to her books, as she believes that each child reads differently ('boekvoorstelling'). While she supports the idea of a minimum age for books, she argues against imposing a maximum age limit and, similarly to Kuijer and Pullman, sees little distinction in her writing for children or adults. However, the analyses reveal that there are discernible differences in certain aspects of her books intended for different readers. Primarily, a clear distinction emerges in terms of style as examined through word and sentence length, as well as lexical diversity. Almond acknowledges a stylistic shift between his short stories for adults and *Skellig*, his first book for 12-year-olds, suggesting a distinction in writing style based on the age of the intended readership (*Interview with David Almond*). The analysis reveals age-related clusters in Almond's oeuvre, although not absolute, indicating that additional factors influence his writing style. Books for 12-year-olds and 15-year-olds show the most stylistic distinction, although some samples overlap with other age ranges.

In conclusion, the digital analyses conducted in this dissertation have provided valuable insights into age-related questions in fiction, specifically regarding the intended reader and the expression of explicit and implicit age ideologies. The research objective of compiling a set of methods for digital text analysis has been achieved, albeit in a non-exhaustive manner. The analyses successfully revealed distinctions between texts intended for readers of different ages, demonstrating the effectiveness of the employed methodologies. It is noteworthy that these distinctions were observed even in cases where authors themselves claimed not to differentiate their writing based on age. Furthermore, the analyses encompassed both large and smaller subsets of the corpus, as well as raw unannotated text material and annotated versions of the texts, showcasing the versatility and robustness of the applied approaches. These findings highlight the potential of digital text analysis in unravelling age-related patterns in fiction and contribute to our understanding of how age ideologies are conveyed through literary works.

8.2 Reflections

The findings of the research conducted in this dissertation, which employed digital text analysis tools to investigate the influence of the intended reader's age on the construction of age in fiction

for young readers, carry significant implications. Through the examination of age representation at both the surface and semantic levels of the text, notable differences were discovered that can be attributed to the age of the intended reader. This suggests that age norms are constructed in distinct ways depending on this categorisation, highlighting the importance of considering the diversity within juvenile literature. It becomes evident that a nuanced understanding of the construction of age in books for different age groups is crucial, rather than treating all juvenile fiction as a homogeneous category, which has been the case in previous research comparing texts for adults to those for younger readers. Moreover, given the influence of age norms in fiction on the socialisation of its readership, it becomes increasingly vital to be mindful of the type of messages children of all ages, adolescents, but also adults encounter when engaging with fiction. This study underscores the necessity of critically examining the content and its impact on readers' development.

The research conducted in this dissertation carries with it some limitations, both in terms of the study as a whole and the way it examines the construction of age, and in its use of a primarily digital methodology. The corpus that was compiled for this dissertation demonstrates efforts towards comprehensiveness in terms of publication period, cultural region, author gender, and ethnicity. However, the corpus cannot claim to be representative of narrative fiction overall. Additionally, the study falls short in considering the impact of context and focalisation, as most analyses extract single words from the text and treat descriptions of characters at face value, disregarding the possible influence of narrative agency on readers' interpretations (see Nikolajeva *Rhetoric*, and Rimmon-Kenan). Incorporating a more nuanced examination of focalisation would have enhanced the study's theoretical framework. An onset to this issue is provided in the analysis of metareflections presented in chapter five, where I consider the source of the statement, whether it originates from a fictional character with their own distinct age or from the third-person narrator. Focalisation, however, is not solely dependent on the character speaking but also encompasses the perspective of the narrator. Due to the complex nature of narrative point of view in textual narration, accurately modelling focalisation in digital text analysis poses challenges. A first step in addressing this challenge might be to differentiate between narration and character speech, with additional information about the character in question. Adding this extra layer of information to the other analyses that look at the construction of age *in* the texts and focus more on content words, such as those in chapters six and seven, might provide interesting new insights into intergenerational relationships. This could contribute to answering the question whether there are differences in the way fictional characters of different ages view a specific age category.

Furthermore, using digital methodologies necessitates certain generalisations to be made to study the construction of fictional age on a larger scale. While such methods require objectivity and uniformity in data, this is not always offered by fictional narratives. For example, some narratives might leave out the age of a character or present it in a vague or ambiguous way. Stories which feature non-human characters and fantastical elements also complicate the attribution of age. Additionally, certain analyses excluded short texts and texts written in idiosyncratic language. More established methods for close reading might be better suited to discuss these special cases and ensure that all texts can be analysed. This highlights the importance of complementing digital analyses with close reading approaches, which excel in capturing nuanced narratives and intricate contextual details. The interaction between these two approaches can solve not only problems of uniformity, but also provides the opportunity to conduct more com-

prehensive, in-depth analyses. To study intergenerational relationships in literature, for example, computational methods, such as sentiment analysis and topic modelling, can be employed to identify patterns in the language and emotions associated with different characters throughout the text. This can provide a quantitative understanding of how characters' relationships evolve over the course of the narrative. However, close reading methods offer a crucial addition to this analysis. Through close reading, specific passages can be studied, specifically the subtleties of language, symbolism, and context. This qualitative approach allows for a nuanced interpretation of character dynamics, motivations, and intentions that may not be captured by computational methods alone.

When looking at each digital tool individually, some additional and more detailed limitations can be identified. Stylometric methods were created primarily to study adult literature and typically rely on linguistic patterns and features that are more prevalent in that category. Literature for younger readers may have distinct stylistic elements and language usage that are age-appropriate and differ from those found in adult texts. Stylometric models may not adequately capture these unique features, leading to potential inaccuracies in the analysis such as a clearer clustering of titles for intended for an adult readership. Surface features, including sentence length and lexical diversity, have the potential to be intentionally manipulated by authors or altered by editors and publishers in order to align with the intended reader's age. Consequently, the findings derived from these analyses could possibly reflect the influence of publisher guidelines to a greater extent than the author's underlying, subconscious choices. The use of the syntactic parser in chapters six and seven was found to be potentially problematic in its reliance on lists of most common words to identify main trends in the description of characters, as this may overlook significant topics described using a more diverse vocabulary. Nevertheless, these limitations do not prompt me to agree with critiques such as those presented by Nan Z. Da. In my study on the construction of age in literature for young readers, I found that digital text analysis tools offer significant advantages, such as the ability to process a large number of texts simultaneously and the identification of patterns and trends across texts that might otherwise go unnoticed. This is especially the case for children's literature, as large-scaled studies in this field of research are still rare. However, it is important to acknowledge that there are instances throughout my research where close reading and critical interpretation remain indispensable for comprehending outliers in the analyses. Hence, I respectfully disagree with Da's assertion that computational literary studies are an unequivocal failure, yet I concur with her cautionary note against the unquestioning adoption of computational approaches.

As one of the primary research objectives of this dissertation was to develop a non-exhaustive set of methods for the study of age-related questions in fiction for young readers, it is important to acknowledge that there are alternative approaches available for examining the construction of age in texts intended for readers of different ages. The existing corpus, comprising 738 unannotated texts of which 238 were also annotated, offers ample potential for conducting additional analyses beyond the scope of this study. One avenue for exploration involves using the annotated material to construct character networks, which can shed light on intergenerational relationships and potential variations in these dynamics based on the age of the intended reader. Furthermore, an in-depth analysis of characters' proper names could contribute to the study of characterisation, considering that such names are included in the character lists. According to Nikolajeva, proper names in children's literature serve multiple functions in terms of characterisation, encompassing their nominative role (*Rhetoric*, 268–70).

The unannotated material also holds significant potential for further investigation, as many surface features have remained unexplored. For instance, functional diversity, a concept that has been extensively studied in ecology research and recently adopted in linguistics and literary studies as an alternative to lexical diversity (see Karsdorp et al.), could be examined within this context.

One key strength of this study lies in its unique methodology, employing digital text analysis tools to explore both explicit and implicit age ideologies within fiction for young readers and adults. Specifically, the examination of implicit age norms has a lot to gain from this methodology. As Stephens suggests, these implicit norms can have a powerful influence on readers, precisely because they are not explicitly addressed or subjected to scrutiny (*Language*, 10). Throughout this dissertation, the analyses yield compelling results that underscore the significance of investigating the construction of age at both the level of the intended reader and the level of fictional age. These findings shed light on the advantages of incorporating computational techniques into the study of fiction intended for readers of different ages, revealing trends, comparing writing styles, and examining the speech of characters across a broad range of texts. One facet of this study centred on examining the attribute of age as it pertains to fictional characters. However, chapter seven also demonstrated the potential of the methodology to explore the interaction between age and other identity characteristics. By employing this approach, future studies might shed light on the intricate dynamics and intersections between age and various dimensions of identity within literature for young readers. By leveraging computational methods, this dissertation also contributes to the existing body of research on crosswriters, expanding the scope of analysis to encompass a greater number of texts and enabling comparative investigations into specific aspects such as the attributes and speech of fictional characters of various ages. While the distant approach employed in this dissertation provides valuable insights, it is important to acknowledge the complementary role of close reading in studying the representation of age in fiction. Many of the results obtained from the analyses can serve as hypotheses to be further examined through close reading. By fostering a fruitful debate between these research traditions, I hope that this study will encourage interdisciplinary approaches within the field of children's literature studies.

Bibliography

Primary works

Aerts, Jef. *De blauwe vleugels*. Querido, 2018.

---. *De nadagen*. De Bezige Bij, 2003.

---. *Haeren Majesteit*. Manteau, 1999.

---. *Het kleine paradijs*. Querido, 2012.

---. *Paard met laarzen*. Querido, 2015.

---. *Rue Fontaine d'Amour*. De Bezige Bij, 2008.

---. *Vertezucht*. Manteau, 2001.

---. *Vissen smelten niet*. Querido, 2013.

Almond, David. *A Song For Ella Grey*. Hodder Children's Books, 2014.

---. *Bone Music*. Hodder Children's Books, 2021.

---. *Brand New Boy*. Walker Books, 2020.

---. *Clay*. Hodder Children's Books, 2005.

---. *Counting Stars*. Hodder Children's Books, 2000.

---. *Harry Miller's Run*. Walker Books, 2015.

---. *Heaven Eyes*. Hodder Children's Books, 2000.

---. *Jackdaw Summer*. Hodder Children's Books, 2008.

---. *Kit's Wilderness*. Hodder Children's Books, 1999.

---. *Klaus Vogel and the Bad Lads*. Barrington Stoke, 2009.

---. *My Dad's a Birdman*. Walker Books, 2007.

---. *My Name is Mina*. Hodder Children's Books, 2010.

---. *Secret Heart*. Hodder Children's Books, 2001.

---. *Skellig*. Hodder Children's Books, 1998.

---. *The Boy Who Climbed into the Moon*. Walker Books, 2010.

---. *The Boy Who Swam With Piranhas*. Walker Books, 2012.

---. *The Colour of the Sun*. Hodder Children's Books, 2018.

---. *The Fire-Eaters*. Hodder Children's Books, 2003.

---. *The Tale of Angelino Brown*. Walker Books, 2017.

---. *The Tightrope Walkers*. Penguin, 2014.

---. *War is Over*. Hodder Children's Books, 2018.

Blackman, Malorie. *A Dangerous Game*. Barrington Stoke, 2018.

---. *An Eye for an Eye*. Corgi Children's, 2003.

---. *A.N.T.I.D.O.T.E*. Doubleday, 1996.

---. *Betsey Biggalow is Here!*. Piccadilly Press, 1992.

---. *Betsey Biggalow the Detective*. Piccadilly Press, 1992.

---. *Betsey's Birthday Surprise*. Piccadilly Press, 1996.

---. *Boys Don't Cry*. Doubleday, 2010.

---. *Callum*. Random House, 2012.

---. *Chasing the Stars*. Random House, 2016.

---. *Checkmate*. Doubleday, 2005.

---. *Cloud Busting*. Yearling, 2004.

---. *Contact*. Barrington Stoke, 2015.

---. *Crossfire*. Random House, 2019

---. *Dangerous Reality*. Doubleday, 1999.

---. *Dead Gorgeous*. Doubleday, 2002.

---. *Double Cross*. Corgi Children's, 2008.

---. *Elaine You're a Brat*. Orchard Books, 1991.

---. *Fangs*. Orchard Books, 1998.

---. *Forbidden Game*. Puffin, 1999.

---. *Girl Wonder and the Terrific Twins*. Penguin, 1991.

---. *Girl Wonder's Winter Adventures*. Puffin, 1992.

---. *Girl Wonder to the Rescue*. Puffin, 1994.

---. *Grandma Gertie's Haunted Handbag*. Heinemann, 1996.

---. *Grandpa Bert and the Ghost Snatchers*. Barrington Stoke, 2018.

---. *Hacker*. Doubleday, 1992.

---. *Hostage*. Barrington Stoke, 1999.

---. *Hurricane Betsey*. Piccadilly Press, 1993.

---. *Jack Sweettooth*. Viking, 1995.

---. *Jon for Short*. Barrington Stoke, 2013.

---. *Knife Edge*. Doubleday, 2004.

---. *Magic Betsey*. Piccadilly Press, 1994.

---. *My Friend's a Gris-Quok*. Scholastic, 1994.

---. *Noble Conflict*. Doubleday, 2013.

---. *Nought Forever*. Puffin, 2019.

---. *Noughts & Crosses*. Doubleday, 2001.

---. *Operation Gadgetman!*. Doubleday, 1993.

---. *Peacemaker*. Heinemann, 2016.

---. *Peril on Planet Pellia*. Orchard Books, 1996.

---. *Pig Heart Boy*. Doubleday, 1997.

---. *Rachel and the Difference Thief*. Longman, 1994.

---. *Rachel versus Bonecrusher the Mighty*. Longman, 1994.

---. *Robot Girl*. Barrington Stoke, 2015.

---. *Sinclair Wonder Bear*. Egmont, 2003.

---. *Snow Dog*. Random House, 2001.

---. *Space Race*. Random House, 1997.

---. *Tell Me No Lies*. Macmillan, 1999.

---. *The Amazing Adventures of Girl Wonder*. Barn Owl, 2003.

---. *The Deadly Dare Mysteries*. Random House, 2005.

---. *The Mellion Moon Mystery*. Orchard Books, 1996.

---. *The Monster Crisp-Guzzler*. Corgi Children's, 2002.

---. *The Quasar Quartz Quest*. Orchard Books, 1996.

---. *The Secret of the Terrible Hand*. Orchard Books, 1996.

---. *The Space Stowaway*. Ginn, 1995.

---. *The Stuff of Nightmare*. Doubleday, 2007.

---. *Thief!*. Doubleday, 1996.

---. *Trust Me*. Women's Press, 1992.

---. *Whizziwig*. Puffin, 1995.

---. *Whizziwig Returns*. Puffin, 1999.

Chambers, Aidan. *Blindside*. Barrington Stoke, 2015

---. *Breaktime*. Bodley Head, 1978.

---. *Cycle Smash*. Heinemann, 1967.

---. *Dance On My Grave*. Bodley Head, 1982.

---. *Dying to Know You*. Bodley Head, 2012.

---. *Marle*. Heinemann, 1968.

---. *Now I Know*. Bodley Head, 1987.

---. *Postcards From No Man's Land*. Bodley Head, 1999.

---. *Seal Secret*. Bodley Head, 1980.

---. *Snake River*. Macmillan, 1975.

---. *The Kissing Game*. Amulet Books, 2011.

---. *The Present Takers*. Bodley Head, 1983.

---. *The Toll Bridge*. Bodley Head, 1992.

---. *This is All*. Bodley Head, 2005.

Cottrell-Boyce, Frank. *Chitty Chitty Bang Bang and the Race Against Time*. Macmillan, 2012.

---. *Chitty Chitty Bang Bang Flies Again*. Macmillan, 2011.

---. *Chitty Chitty Bang Bang Over the Moon*. Macmillan, 2013.

---. *Cosmic*. Macmillan, 2008.

---. *Desirable*. Barrington Stoke, 2008.

---. *Framed*. Macmillan, 2005.

---. *Millions*. HarperCollins, 2004.

---. *Runaway Robot*. Macmillan, 2019.

---. *Sputnik's Guide to Life on Earth*. Macmillan, 2016.

---. *Ted Rules the World*. Barrington Stoke, 2015.

---. *The Astounding Broccoli Boy*. Macmillan, 2015.

---. *The Great Rocket Robbery*. Macmillan, 2019.

---. *The Unforgotten Coat*. Walker Books, 2011.

Dumon Tak, Bibi. *Bezem*. Van Goor, 2005.

---. *De dag dat ik mijn naam veranderde*. De Geus, 2020.

---. *De eik was hier*. Querido, 2021.

---. *Laïka tussen de sterren*. Querido, 2006.

---. *Latino King*. Querido, 2010.

---. *Mikis de ezeljongen*. Querido, 2011.

---. *Rotjongens*. Athenaeum-Polak & Van Gennep, 2007.

---. *Soldaat Wojtek*. Querido, 2008.

---. *Wolfskwint*. Athenaeum-Polak & Van Gennep, 2013.

Fine, Anne. *All Bones and Lies*. Black Swan, 2001.

---. *Anneli the Art Hater*. Egmont, 1986.

---. *A Pack of Liars*. Puffin, 1988.

---. *A Sudden Glow of Gold*. Piccadilly Press, 1991.

---. *A Sudden Puff of Glittering Smoke*. Piccadilly Press, 1989.

---. *A Sudden Swirl of Icy Wind*. Piccadilly Press, 1990.

---. *Back on Weird Street*. Barrington Stoke, 2018.

---. *Bad Dreams*. Doubleday, 2000.

---. *Battle of Wills*. Endeavour, 2016.

---. *Bill's New Frock*. Methuen, 1989.

---. *Blood Family*. Doubleday, 2013.

---. *Blue Moon Day*. Corgi Children's, 2014.

---. *Care of Henry*. Walker Books, 1996.

---. *Charm School*. Doubleday, 1999.

---. *Countdown*. Heinemann, 1996.

---. *Crummy Mummy and Me*. Puffin, 1988.

---. *Design a Pram*. Egmont, 1991.

---. *Eating Things on Sticks*. Doubleday, 2009.

---. *Flour Babies*. Hamish Hamilton, 1992.

---. *Fly in the Ointment*. Bantam Press, 2008.

---. *Friday Surprise*. Egmont, 2010.

---. *Frozen Billy*. Doubleday, 2004.

---. *Gnomes, Gnomes, Gnomes*. Barrington Stoke Ltd, 2013.

---. *Goggle-Eyes*. Hamish Hamilton, 1989.

---. *Hole in the Road*. Collins Educational, 2014.

---. *How Brave Is That?*. Barrington Stoke Ltd, 2013.

---. *How to Cross the Road and Not Turn into a Pizza*. Walker Books, 2002.

---. *How To Write Really Badly*. Methuen, 1996.

---. *In Cold Domain*. Viking, 1994.

---. *Into the Bin*. Barrington Stoke Ltd, 2019.

---. *It Moved!*. Walker Books, 2006.

---. *Ivan the Terrible*. Egmont, 2007.

---. *Jamie and Angus Forever*. Walker Books, 2009.

---. *Jamie and Angus Together*. Walker Books, 2007.

---. *Jennifer's Diary*. Hamish Hamilton, 1996.

---. *Loudmouth Louis*. Penguin, 1998.

---. *Madame Doubtfire*. Hamish Hamilton, 1987.

---. *Magic Ball*. Egmont, 2017.

---. *Notso Hotso*. Hamish Hamilton, 2001.

---. *Only a Show*. Hamish Hamilton, 1990.

---. *On Planet Fruitcake*. Egmont, 2013.

---. *Our Precious Lulu*. Bantam Press, 2009.

---. *Press Play*. Piccadilly Press, 1994.

---. *Raking the Ashes*. Bantam Press, 2005.

---. *Roll Over Roly*. Puffin, 1999.

---. *Round Behind the Ice-House*. Methuen, 1981.

---. *Saving Miss Mirabelle*. Egmont, 2002.

---. *Scaredy-Cat*. Heinemann, 1985.

---. *Step by Wicked Step*. Hamish Hamilton, 1995.

---. *Stranger Danger?*. Hamish Hamilton, 1989.

---. *Taking the Devil's Advice*. Viking, 1990.

---. *Tales from Weird Street*. Barrington Stoke Ltd, 2017.

---. *Telling Liddy*. Bantam Press, 1998.

---. *The Angel of Nitshill Road*. Methuen, 1992.

---. *The Book of the Banshee*. Corgi Children's, 1991.

---. *The Chicken Gave it to Me*. Methuen, 1992.

---. *The Devil Walks*. Doubleday, 2011.

---. *The Diary of a Killer Cat*. Hamish Hamilton, 1994.

---. *The Gingerbread Star*. Barrington Stoke Ltd, 2015.

---. *The Granny Project*. Farrar Straus and Giroux, 1983.

---. *The Haunting of Pip Parker*. Walker Books, 1991.

---. *The Haunting of Uncle Ron*. Barrington Stoke Ltd, 2014.

---. *The Jamie and Angus Stories*. Walker Books, 2002.

---. *The Killer Cat Runs Away*. Doubleday, 2013.

---. *The Killer Cat's Birthday Bash*. Puffin, 2008.

---. *The Killer Cat's Christmas*. Puffin, 2009.

---. *The Killer Cat Strikes Back*. Puffin, 2007.

---. *The Killjoy*. Bantam Press, 1986.

---. *The More the Merrier*. Doubleday, 2003.

---. *The Nag Club*. Walker Books, 2004.

---. *The Other Darker Ned*. Methuen, 1979.

---. *The Party Club*. Walker Books, 2015.

---. *The Return of the Killer Cat*. Puffin, 2003.

---. *The Road of Bones*. Corgi Children's, 2006.

---. *The Same Old Story Every Year*. Hamish Hamilton, 1992.

---. *The Stone Menagerie*. Corgi Children's, 1980.

---. *The Summer-House Loon*. Methuen, 1978.

---. *The Tulip Touch*. Puffin, 1996.

---. *The Worst Child I Ever Had*. Hamish Hamilton, 1991.

---. *Trouble in Toadpool*. Doubleday, 2012.

---. *Under a Silver Moon*. Walker Books, 2010.

---. *Under the Bed*. Egmont, 2015.

---. *Up on Cloud Nine*. Doubleday, 2002.

---. *Very Different*. Mammoth, 2001.

Franck, Ed. *Amerika of de Noordpool*. Clavis, 1990.

---. *Begraaf me over de bergen*. Houtekiet, 1988.

---. *Dazyl en de twee stenen beelden*. Davidsfonds, 2016.

---. *De dolk*. Clavis, 1995.

---. *De droge rivier*. Altiora, 1997.

---. *De duistere stem van het eiland*. Clavis, 1992.

---. *De hemel bestaat, uit vijf letters*. Clavis, 1993.

---. *De koningen van de straat*. Standaard, 1997.

---. *De liefste of de dapperste of allebei*. Davidsfonds, 2006.

---. *De man bij het meer*. Querido, 2005.

---. *De verrijzenis van Tony Tornado*. Clavis, 1992.

---. *De witte muur*. Davidsfonds, 1988.

---. *Diederik de Derde*. Altiora, 2000.

---. *Diederik Piepkuiken*. Clavis, 1991.

---. *Een kanarie in mijn hoofd*. Davidsfonds, 2010.

---. *Een panda voor Wimpie*. Artis-Historia, 1999.

---. *Een vreemde stoet*. Altiora, 1995.

---. *Els heeft een pels*. Altiora, 1995.

---. *Geen wonder dat moeder met de goudvissen praat*. Clavis, 1988.

---. *Het glazen masker*. Clavis, 1992.

---. *Het huis van eb en vloed*. Altiora, 1998.

---. *Hoe Seppe ondanks alles zijn zestiende verjaardag haalde*. Davidsfonds, 2012.

---. *In de greep van de macht*. Davidsfonds, 1989.

---. *Jelle en Jolien*. Altiora, 1998.

---. *Jolientje fantast*. Van Halewyck, 2018.

---. *Manke Bever*. Davidsfonds, 1987.

---. *Met de kont tegen de krib*. Clavis, 1989.
---. *Mijn zus draagt een heuvel op haar rug*. Altiora, 2000.
---. *Minne, het mozaïekmeisje*. Davidsfonds, 2015.
---. *Moet je echt weg?*. Clavis, 1987.
---. *Spetters op de kermis*. Clavis, 1985.
---. *Tenten in de wei*. Clavis, 1986.
---. *Tijger*. Altiora, 1995.
---. *Tinka*. Altiora, 2001.
---. *Tony Tornado en de linkerborst*. Clavis, 1994.
---. *Uit de loop van een geweer*. Clavis, 1991.
---. *Vechten om een levend huis*. Davidsfonds, 1989.
---. *Wie wil Wubbe weg?*. Clavis, 1989.
---. *Zie ik je nog eens terug?*. Querido, 2003.
---. *Zomer Zeventien*. Clavis, 1990.
Gaiman, Neil. *American Gods*. Headline, 2001.
---. *Anansi Boys*. HarperCollins, 2005.
---. *Coraline*. HarperCollins, 2002.
---. *Eternity's Wheel*. HarperCollins, 2015.
---. *Fortunately, the Milk*. HarperCollins, 2013.
---. *Good Omens*. Gollancz, 1990.
---. *How the Marquis Got His Coat Back*. Headline, 2015.
---. *InterWorld*. HarperCollins, 2007.
---. *M is for Magic*. HarperCollins, 2007.
---. *Neverwhere*. BBC Books, 1996.
---. *Odd and the Frost Giants*. HarperCollins, 2008.
---. *Stardust*. Headline, 1999.
---. *The Graveyard Book*. HarperCollins, 2008.
---. *The Ocean at the End of the Lane*. HarperCollins, 2013.
---. *The Silver Dream*. HarperCollins, 2013.
---. *The Sleeper and the Spindle*. Bloomsbury, 2013.
---. *The Truth is a Cave in the Black Mountains*. HarperCollins, 2014.
Gavin, Jamila. *Ali and the Robots*. Methuen, 1986.
---. *Blackberry Blue*. Tamarind Books, 2013.
---. *Coram Boy*. Egmont, 2000.

---. *Deadly Friend*. Heinemann, 1994.

---. *Digital Dan*. Methuen, 1984.

---. *Double Dare*. Methuen, 1982.

---. *Fine Feathered Friend*. Heinemann, 1996.

---. *Forbidden Memories*. Mammoth, 1998.

---. *From out of the Shadows*. Egmont, 2002.

---. *Grandma's Surprise*. Heinemann, 1996.

---. *Grandpa Chatterji's Third Eye*. Egmont, 2006.

---. *Grandpa's Indian Summer*. Methuen, 1995.

---. *I Want to be an Angel*. Methuen, 1990.

---. *Kamla and Kate*. Methuen, 1983.

---. *Kamla and Kate Again*. Methuen, 1991.

---. *Out of India*. Pavilion, 1997.

---. *Presents*. Heinemann, 1996.

---. *Someone's Watching, Someone's Waiting*. Mammoth, 1998.

---. *Starchild on Clark Street*. Cambridge University Press, 1998.

---. *The Eye of the Horse*. Methuen, 1994.

---. *The Hideaway*. Methuen, 1987.

---. *The Mango Tree*. Heinemann, 1996.

---. *The Robber Baron's Daughter*. Egmont, 2008.

---. *The Singing Bowls*. Methuen, 1989.

---. *The Track of the Wind*. Mammoth, 1997.

---. *The Wheel of Surya*. Methuen, 1992.

---. *The Wormholers*. Mammoth, 1996.

---. *Walking on my Hands*. Hodder Children's Books, 2007.

---. *Who Did It?*. Heinemann, 1996.

Heesen, Martha. *Bajaar*. Querido, 2011.

---. *Biezel*. Querido, 2014.

---. *De ijzeren hemel*. Querido, 2004.

---. *De laatste jongen*. Querido, 1995.

---. *De vloek van Cornelia*. Querido, 1999.

---. *De vreemde vrouw van de lange straat*. Querido, 1996.

---. *Een held van klei*. Querido, 1994.

---. *Het plan Stoffel*. Querido, 1993.

---. *Maandag heeft vleugels*. Querido, 2005.

---. *Mijn broer, de nieuwe en ik*. Querido, 2009.

---. *Mijn zusje is een monster*. Querido, 2000.

---. *Stekels*. Querido, 2001.

---. *Sterre en Joe*. Querido, 1997.

---. *Toen Faas niet thuis kwam*. Querido, 2003.

---. *Watson*. Querido, 2007.

---. *Wolf*. Querido, 2006.

Hof, Marjolijn. *Als niemand kijkt*. Querido, 2009.

---. *De bloem van de buurt*. NZV, 1999.

---. *De regels van drie*. Querido, 2013.

---. *Een kleine kans*. Querido, 2006.

---. *Het panterfeestje*. NZV, 2000.

---. *Lepelsnijder*. Querido, 2018.

---. *Mijn opa en ik en het varken oma*. Querido, 2011.

---. *Moeder nummer nul*. Querido, 2008.

---. *Oversteken*. Querido, 2007.

---. *Toeter*. NZV, 2001.

Kuijer, Guus. *De jonge prinsen*. Querido, 1986.

---. *De man met de hamer*. De Arbeiderspers, 1975.

---. *De redder van Afrika*. De Arbeiderspers, 1989.

---. *De tranen knallen uit mijn kop*. Querido, 1980.

---. *De zwarte stenen*. Querido, 1984.

---. *Drie verschrikkelijke dagen*. Querido, 1976.

---. *Eend voor eend*. Querido, 1983.

---. *Een gat in de grens*. Querido, 1975.

---. *Een hoofd vol macaroni*. Querido, 1979.

---. *Florian Knol*. Querido, 2006.

---. *Grote mensen, daar kan je beter soep van koken*. Querido, 1976.

---. *Het boek van alle dingen*. Querido, 2004.

---. *Het dochtertje van de wasvrouw*. Querido, 1980.

---. *Het geluk komt als de donder*. Querido, 2000.

---. *Het is fijn om er te zijn*. Querido, 2000.

---. *Het land van de neushoornvogel*. Querido, 1985.

---. *Het vogeltje van Amsterdam*. De Arbeiderspers, 1992.

---. *Hoe Mieke Mom haar maffe moeder vindt*. Querido, 1978.

---. *Ik ben Polleke, hoor!*. Querido, 2001.

---. *Ik woonde in een leunstoel*. Querido, 1979.

---. *Izebel van Tyrus*. De Arbeiderspers, 1988.

---. *Krassen in het tafelblad*. Querido, 1978.

---. *Met de poppen gooien*. Querido, 1975.

---. *Met de wind mee naar de zee*. Querido, 2001.

---. *Olle*. Querido, 1990.

---. *Op je kop in de prullenbak*. Querido, 1977.

---. *Pappa is een hond*. Querido, 1977.

---. *Rose, met vrome wimpers*. Meulenhoff, 1971.

---. *Tin Toeval en de kunst van het verdwalen*. Querido, 1987.

---. *Tin Toeval en de kunst van Madelief*. Querido, 1989.

---. *Tin Toeval en het geheim van Tweebeens-eiland*. Querido, 1987.

---. *Tin Toeval in de onderwereld*. Querido, 1993.

---. *Voor altijd samen, amen*. Querido, 1999.

Moeyaert, Bart. *Afrika achter het hek*. Querido, 1995.

---. *Blote handen*. Querido, 1995.

---. *Broere*. Querido, 2000.

---. *Dani Bennoni*. Querido, 2004.

---. *De baas van alles*. Manteau, 2007.

---. *De brief die Rosie vond*. Standaard, 1997.

---. *De gans en zijn broer*. Querido, 2014.

---. *De koning en de koffieboon*. De Fontein, 1998.

---. *De lente van '39*. Ezelsoor, 1988.

---. *De man in de maan*. Zwijsen, 1992.

---. *De melkweg*. Querido, 2011.

---. *De zeug, de beer en de big*. Dietsche Warande & Belfort, 2002.

---. *Die steeg van ons*. Zwijsen, 1995.

---. *Duet met valse noten*. Altiora, 1983.

---. *Echt weg is niet zo ver*. Zwijsen, 1993.

---. *Een boot in de mist*. Zwijsen, 1990.

---. *Een klap is geen kus*. Zwijsen, 1989.

---. *Een kuil om in te wonen*. Zwijsen, 1990.

---. *Eerst zien*. Nationaal Centrum voor Jeugdliteratuur, 1997.

---. *Graz*. Querido, 2009.

---. *Het beest heet Mona*. Zwijsen, 2001.

---. *Het boek van Niete*. Altiora, 1988.

---. *Het is de liefde die we niet begrijpen*. Querido, 1999.

---. *Hij was de haas*. Nationaal Centrum voor Jeugdliteratuur, 1996.

---. *Iemands lief*. 2013.

---. *Ik woon in een ruimteschip*. 2013.

---. *Kus me*. Altiora, 1991.

---. *Lamelos*. 2010.

---. *Mansoor*. Querido, 1996.

---. *Midwinternachtsdroom*. 1991.

---. *Mijn tuin uit!*. Zwijsen, 1989.

---. *Negen keer Nina*. 1990.

---. *Parkietje*. 1989.

---. *Suzanne Dantine*. Altiora, 1989.

---. *Tegenwoordig heet iedereen Sorry*. Querido, 2018.

---. *Terug naar af*. Altiora, 1986.

---. *Tik tik*. Bakermat, 1992.

---. *Voor altijd, altijd*. Zwijsen, 1992.

---. *Wakker kussen*. Davidsfonds, 1988.

---. *Wespennest*. Querido, 1997.

Murphy, Jill. *A Bad Spell for the Worst Witch*. Kestrel Books, 1982.

---. *First Prize for the Worst Witch*. Puffin, 2018.

---. *Jeffrey Strangeways*. Walker Books, 1992.

---. *The Worst Witch*. Alison & Busby, 1974.

---. *The Worst Witch All at Sea*. Viking, 1993.

---. *The Worst Witch and the Wishing Star*. Puffin, 2013.

---. *The Worst Witch Saves the Day*. Puffin, 2005.

---. *The Worst Witch Strikes Again*. Alison & Busby, 1980.

---. *The Worst Witch to the Rescue*. Puffin, 2007.

---. *Worlds Apart*. Walker Books, 1988.

Provoost, Anne. *De arkvaarders*. Querido, 2001.

---. *De camcorder*. Averbode, 1994.

---. *De roos en het zwijn*. Querido, 1997.

---. *De wekker en het mes*. Zwijsen, 1991.

---. *Het horloge*. Altiora, 1992.

---. *In de zon kijken*. Querido, 2007.

---. *Kauwgom voor de held*. Zwijsen, 1993.

---. *Luchtpost*. 1987.

---. *Mijn tante is een grindewal*. Houtekiet, 1990.

---. *Niet uitlachen!*. Zwijsen, 1991.

---. *Vallen*. Houtekiet, 1994.

Pullman, Philip. *Clockwork*. Doubleday, 1996.

---. *Count Karlstein*. Chatto & Windus, 1982.

---. *Galatea*. Littlehampton Book Services, 1978.

---. *How to be Cool*. Heinemann, 1987.

---. *I was a Rat!*. Doubleday, 1999.

---. *La Belle Sauvage*. David Fickling Books, 2017.

---. *Lyra's Oxford*. David Fickling Books, 2003.

---. *Mossycoat*. Scholastic, 1998.

---. *Northern Lights*. Scholastic Children's Books, 1995.

---. *Once Upon a Time in the North*. David Fickling Books, 2008.

---. *Serpentine*. Penguin Random House, 2020.

---. *Spring-Heeled Jack*. Doubleday, 1989.

---. *The Amber Spyglass*. David Fickling Books, 2000.

---. *The Broken Bridge*. Macmillan Children's Books, 1990.

---. *The Collectors*. RHCP Digital, 2014.

---. *The Firework-Maker's Daughter*. Doubleday, 1995.

---. *The Gas-Fitter's Ball*. Viking, 1995.

---. *The Good Man Jesus and the Scoundrel Christ*. Canongate, 2010.

---. *The Haunted Storm*. New English Library, 1972.

---. *The Ruby in the Smoke*. Oxford University Press, 1985.

---. *The Scarecrow and his Servant*. Doubleday, 2004.

---. *The Secret Commonwealth*. David Fickling Books, 2019.

---. *The Shadow in the North*. Oxford University Press, 1986.

---. *The Subtle Knife*. Scholastic, 1997.

---. *The Tiger in the Well*. Scholastic, 1990.

---. *The Tin Princess*. Penguin, 1994.

---. *The White Mercedes*. Macmillan Children's Books, 1992.

---. *Thunderbolt's Waxwork*. Viking, 1994.

Remmerts de Vries, Daan. *Bernie King en de magische cirkels*. Querido, 2008.

---. *Bernie King in een donkere spiegel*. Querido, 2009.

---. *Brave nieuwe wereld*. Querido, 2011.

---. *Circus pingies een avontuur van Zippy en Slos*. Querido, 2006.

---. *De blauwe maanvis*. Querido, 2016.

---. *De diepte van een Zweeds meer*. Lemniscaat, 2006.

---. *De harpij*. Querido, 2014.

---. *De kerstnacht van Zippy en Slos*. Hoogland & Van Klaveren, 2013.

---. *De liefste Olivier*. Leopold, 2004.

---. *De nadagen*. Querido, 2018.

---. *De noordenwindheks*. Querido, 2004.

---. *De ontdekkingsreizigers*. Leopold, 1997.

---. *De stille oorlog van Olivier*. Leopold, 2001.

---. *Een wild feest*. Zwijsen, 1999.

---. *Geest*. Querido, 2017.

---. *Godje*. Querido, 2002.

---. *Groter dan de lucht, erger dan de zon*. Querido, 2015.

---. *Het leven achter de dingen*. Hoogland & Van Klaveren, 2018.

---. *Ijspaleis*. Prometheus, 2003.

---. *Magic Olivier*. Leopold, 1993.

---. *Olivier en het beest*. Leopold, 2003.

---. *Oliviers dagboek*. Elzenga, 2001.

---. *Over het uitkomen van wensen*. Kinderboekenmarkt, 2005.

---. *Tijgereiland*. Querido, 2013.

---. *Wie is Libby Skibner*. Querido, 2009.

---. *Zippy en Slos*. Altamira, 1990.

---. *Zo zal ik bij je zijn*. Prometheus, 2006.

Rowling, J.K.. *Harry Potter and the Chamber of Secrets*. Bloomsbury, 1998.

---. *Harry Potter and the Deathly Hallows*. Bloomsbury, 2007.

---. *Harry Potter and the Goblet of Fire*. Bloomsbury, 2000.

---. *Harry Potter and the Half-Blood Prince*. Bloomsbury, 2005.

---. *Harry Potter and the Order of the Phoenix*. Bloomsbury, 2003.

---. *Harry Potter and the Philosopher's Stone*. Bloomsbury, 1997.

---. *Harry Potter and the Prisoner of Azkaban*. Bloomsbury, 1999.

---. *The Casual Vacancy*. Little Brown and Company, 2012.

---. *The Ickabog*. Hachette, 2020.

Rowling, J.K. (penname: Robert Galbraith). *Career of Evil*. Sphere, 2015.

---. *Lethal White*. Sphere, 2018.

---. *The Cuckoo's Calling*. Sphere, 2013.

---. *The Silkworm*. Sphere, 2014.

---. *Troubled Blood*. Sphere, 2020.

Sax, Aline. *De dodendraad*. Averbode, 2014.

---. *De gebroken harp*. Clavis, 2004.

---. *De hond van Roosevelt*. Averbode, 2007.

---. *De jongen op het dak*. De Eenhoorn, 2020.

---. *De kleuren van het ghetto*. De Eenhoorn, 2011.

---. *De laatste reis*. Clavis, 2010.

---. *De lantaarnaansteker*. De Eenhoorn, 2019.

---. *Duivelsvlucht*. Clavis, 2002.

---. *Een dodelijk bericht*. Averbode, 2013.

---. *Geen stap terug*. Clavis, 2005.

---. *Grensgangers*. Davidsfonds, 2015.

---. *Het meisje en de soldaat*. De Eenhoorn, 2013.

---. *In de schaduw van de zwarte dood*. Averbode, 2010.

---. *Mist over het strand*. Clavis, 2001.

---. *Schaduwleven*. Clavis, 2007.

---. *Wij, twee jongens*. Clavis, 2006.

Van der Geest, Simon. *Geel gras*. Querido, 2009.

---. *Het werkstuk*. Querido, 2019.

---. *Per ongelukt!*. Stichting Collectieve Propaganda van het Nederlandse Boek, 2015.

---. *Spijkerzwijgen*. Querido, 2015.

---. *Spinder*. Querido, 2012.

---. *Wie doet er mee met de zombie race?*. Querido, 2021.

---. *Wie is er bang voor de voetbalzombies?*. Querido, 2016.

Vandermeeren, Hilde. *Als alles duister wordt*. Querido, 2013.

---. *Bodemloos*. Van Halewyck, 2019.

---. *Camping Zeevos*. Davidsfonds, 2012.

---. *De club van Sherlock*. Averbode, 2013.

---. *De jacht op de valse koekjes*. Delubas, 2014.

---. *De prinses en de kurk*. Davidsfonds, 2002.

---. *De toeschouwers*. Querido, 2014.

---. *De twaalfde man*. De Eenhoorn, 2004.

---. *De vuurman*. Delubas, 2012.

---. *Een huis om in te verdwalen*. Davidsfonds, 2002.

---. *Een vroege zomer*. Davidsfonds, 2001.

---. *Het gebroken masker*. Davidsfonds, 2003.

---. *Het geheim van Barry Boudini*. Maretak, 2012.

---. *Het geheim van Cornwall Castle*. Davidsfonds, 2010.

---. *Het grote Paulien-plan*. Standaard, 2011.

---. *Het kistje van Cleo*. Davidsfonds, 2005.

---. *Het mysterie van de drie vazen*. Davidsfonds, 2011.

---. *Het mysterie van Ravenheim*. Averbode, 2014.

---. *Het uitzonderlijke leven van Violetta Vonk*. Davidsfonds, 2013.

---. *Maankind*. Davidsfonds, 2009.

---. *Mama wil op reis*. De Eenhoorn, 2008.

---. *Meedogenloos*. Van Halewyck, 2020.

---. *Mijn geheime papa*. Davidsfonds, 2008.

---. *Moord in de wijk*. Vrijdag, 2019.

---. *Nachtraven en de zaak van de verdwenen postbode*. Davidsfonds, 2009.

---. *Operatie Bernie buiten*. Davidsfonds, 2009.

---. *Pas op voor de buren*. Querido, 2018.

---. *Rusteloos*. Van Halewyck, 2018.

---. *Schatten zoeken*. De Eenhoorn, 2007.

---. *Schemerzone*. Querido, 2017.

---. *Scorpio*. Querido, 2016.

---. *Stille grond*. Querido, 2015.

---. *Strijd op de dansvloer*. Standaard, 2012.

---. *Verboden te lachen*. De Eenhoorn, 2012.

---. *Wolken boven Waterdorp*. Davidsfonds, 2015.

Van de Vendel, Edward. *De dagen van de bluegrassliefde*. Querido, 1999.

---. *De gelukvinder*. Querido, 2008.

---. *De grote verboden zolder*. Querido, 2017.

---. *De Hamburger Files*. Averbode, 2007.

---. *De jongen die met de dieren schaatste*. Querido, 2018.

---. *De lol*. Zwijsen, 2004.

---. *De raadsels van Sam*. Querido, 2012.

---. *Dertien rennende hertjes*. De Eenhoorn, 2012.

---. *Dom konijn*. Querido, 2000.

---. *Een verhaal met een tong*. Lemniscaat, 2006.

---. *Gloeiende voeten. Querido*, 2004.

---. *Gooi dit boek maar uit het raam*. Zwijsen, 2005.

---. *Het bamboe meisje*. Querido, 2021.

---. *Het kankerkampioenschap voor junioren*. Querido, 2015.

---. *Het lekkere van pesten*. 2011.

---. *Jaap deelt klappen uit*. Querido, 1999.

---. *Miss Eenhoorn*. Querido, 2019.

---. *Oliver*. Querido, 2015.

---. *Onder schot*. Leopold, 2009.

---. *Ons derde lichaam*. Querido, 2006.

---. *Prins Nul*. Zwijsen, 2004.

---. *Pup en Kit*. De Eenhoorn, 2002.

---. *Sofie en de dolfijnen*. Querido, 2014.

---. *Sofie en de olifanten*. Querido, 2018.

---. *Sofie en de pinguïns*. Querido, 2010.

---. *Sofie en het geheime paard*. Querido, 2015.

---. *Sofie en het ijsbeertje*. Querido, 2013.

---. *Sofie en het vliegende jongetje*. Querido, 2012.

---. *Toen kwam Sam*. Querido, 2011.

---. *Waar kunnen we hier een standbeeld krijgen?*. Querido, 2007.

---. *Wat ik vergat*. Querido, 2001.

---. *Wat rijmt er op puree?*. Querido, 2005.

---. *Wie wil er met ons trouwen?*. Querido, 2008.

---. *Zootje was hier*. De Eenhoorn, 2004.

Van Leeuwen, Joke. *Alles nieuw*. Querido, 2008.

---. *Bezoekjaren*. Querido, 1988.

---. *De Appelmoesstraat is anders*. Querido, 1978.

---. *Deesje*. Querido, 1985.

---. *De metro van Magnus*. Omniboek, 1981.

---. *De onervarenen*. Querido, 2015.

---. *De tjilpmachine*. Querido, 1990.

---. *Dit boek heet anders*. Querido, 1992.

---. *Een huis met zeven kamers*. Omniboek, 1979.

---. *Feest van het begin*. Querido, 2012.

---. *Het verhaal van Bobbel*. Querido, 1987.

---. *Hier*. Querido, 2018.

---. *Iep!*. Querido, 1996.

---. *Ik heet Reinier en ons huis is afgebrand*. Querido, 2020.

---. *Kukel*. Querido, 1998.

---. *Kweenie*. Querido, 2003.

---. *Maar ik ben Frederik, zei Frederik*. Querido, 2013.

---. *Slopie*. Querido, 2004.

---. *Toen ik*. Querido, 2017.

---. *Toen mijn vader een struik werd*. Querido, 2010.

---. *Vrije vormen*. Querido, 2002.

---. *Wijd weg*. Querido, 1991.

Verschuur, Rita. *Achter de bergen ligt de zee*. Querido, 1987.

---. *De driehoeksdans*. Van Goor, 2004.

---. *De prijs*. Van Goor, 2001.

---. *De torenvalk*. Van Goor, 1990.

---. *De tweede moeder*. Cossee, 2011.

---. *Drie is te veel*. Ploegsma, 1982.

---. *Een droom van een oom*. Zwijsen, 1983.

---. *Een rooskleurig meisje*. Ploegsma, 1981.

---. *Gewoon mezelf*. Ploegsma, 1983.

---. *Het geheim van mijn moeder*. Cossee, 2008.

---. *Het lied van de blauwe parkiet*. Ploegsma, 1985.

---. *Het oude huis op het duin*. Van Goor, 1991.

---. *Hoe moet dat nu met de papillotten*. Van Goor, 1993.

---. *Hoofdbagage*. Van Goor, 1996.

---. *Ik en mijn broertje, mijn broertje en ik*. Ploegsma, 1980.

---. *Jubeltenen*. Van Goor, 1998.

---. *Kinderen en gekken*. Ploegsma, 1978.

---. *Met wortel en tak*. Lemniscaat, 2014.

---. *Mijn hersens draaien rondjes*. Van Goor, 1994.

---. *Moeder en God en ik*. Van Goor, 2005.

---. *Vreemd land*. Van Goor, 1995.

---. *Wat een bof*. Zwijsen, 1982.

---. *Wat je zegt ben je zelf*. Ploegsma, 1977.

---. *Wie is hier eigenlijk de baas*. Ploegsma, 1977.

---. *Ze snappen er niets van*. Ploegsma, 1976.

---. *Zwerftocht van een trollen kind*. Ploegsma, 1978.

Wilson, Jacqueline. *Amber*. Oxford University Press, 1986.

---. *Bad Girls*. Corgi, 1996.

---. *Best Friends*. Corgi, 2004.

---. *Big Day Out*. Corgi, 2012.

---. *Buried Alive!*. Corgi, 1998.

---. *Candyfloss*. Corgi, 2006.

---. *Clean Break*. Doubleday, 2005.

---. *Cliffhanger*. Corgi, 1995.

---. *Clover Moon*. Penguin, 2016.

---. *Connie and the Water Babies*. Methuen, 1996.

---. *Cookie*. Doubleday, 2008.

---. *Dancing the Charleston*. Doubleday, 2019.

---. *Deep Blue*. Oxford University Press, 1993.

---. *Diamond*. Doubleday, 2013.

---. *Double Act*. Doubleday, 1995.

---. *Dustbin Baby*. Doubleday, 2001.

---. *Emerald Star*. Doubleday, 2012.

---. *Falling Apart*. Oxford University Press, 1989.

---. *Girls in Love*. Doubleday, 1997.

---. *Girls in Tears*. Doubleday, 2002.

---. *Girls out Late*. Doubleday, 1999.

---. *Girls under Pressure*. Doubleday, 1998.

---. *Glubbslyme*. Oxford University Press, 1987.

---. *Hetty Feather*. Doubleday, 2009.

---. *Hetty Feather's Christmas*. Doubleday, 2017.

---. *Hide and Seek*. Macmillan, 1972.

---. *How to Survive Summer Camp*. Oxford University Press, 1985.

---. *Jacky Daydream*. Corgi, 2007.

---. *Jimmy Jelly*. Piccadilly Press, 1995.

---. *Kiss*. Corgi, 2007.

---. *Let's Pretend*. Macmillan, 1976.

---. *Lily Alone*. Doubleday, 2011.

---. *Little Darlings*. Doubleday, 2010.

---. *Little Stars*. Doubleday, 2015.

---. *Lizzie Zipmouth*. Young Corgi, 2000.

---. *Lola Rose*. Doubleday, 2003.

---. *Love Frankie*. Doubleday, 2020.

---. *Love From Katy*. 1995.

---. *Love Lessons*. Doubleday, 2005.

---. *Making Hate*. Macmillan, 1977.

---. *Mark Spark in the Dark*. Puffin, 1993.

---. *Midnight*. Doubleday, 2003.

---. *Monster Eyeballs*. Egmont, 1999.

---. *Mr. Cool*. Kingfisher, 1996.

---. *My Brother Bernadette*. Egmont, 1995.

---. *My Mum Tracy Beaker*. Penguin, 2018.

---. *My Secret Diary*. Doubleday, 2009.

---. *My Sister Jodie*. Corgi, 2008.

---. *Nobody's Perfect*. Oxford University Press, 1982.

---. *Opal Plumstead*. Doubleday, 2014.

---. *Queenie*. Doubleday, 2013.

---. *Rent A Bridesmaid*. Penguin, 2016.

---. *Rose Rivers*. Penguin, 2018.

---. *Sapphire Battersea*. Doubleday, 2011.

---. *Secrets*. Doubleday, 2002.

---. *Sleepovers*. Doubleday, 2001.

---. *Snap*. Macmillan, 1974.

---. *Sophie's Secret Diary*. Ginn, 1995.

---. *Starring Tracy Beaker*. Doubleday, 2006.

---. *Take a Good Look*. Blackie and Son Ltd, 1990.

---. *The Bed and Breakfast Star*. Doubleday, 1994.

---. *The Butterfly Club*. Doubleday, 2015.

---. *The Cat Mummy*. Corgi, 2001.

---. *The Dare Game*. Doubleday, 2000.

---. *The Diamond Girls*. Doubleday, 2004.

---. *The Dinosaur's Packed Lunch*. Doubleday, 1995.

---. *The Dream Palace*. Oxford University Press, 1991.

---. *The Illustrated Mum*. Doubleday, 1999.

---. *The Killer Tadpole*. Hamish Hamilton, 1984.

---. *The Left-Outs*. Blackie, 1989.

---. *The Longest Whale Song*. Doubleday, 2010.

---. *The Lottie Project*. Doubleday, 1997.

---. *The Monster in the Cupboard*. Blackie, 1986.

---. *The Monster Story-Teller*. Doubleday, 1997.

---. *The Mum-Minder*. Yearling, 1993.

---. *The Other Side*. Oxford University Press, 1984.

---. *The Party in the Lift*. Blackie, 1989.

---. *The Power of the Shade*. Oxford University Press, 1987.

---. *The School Trip*. Hamish Hamilton, 1984.

---. *The Story of Tracy Beaker*. Doubleday, 1991.

---. *The Suitcase Kid*. Yearling, 1992.

---. *The Worry Website*. Yearling, 2002.

---. *The Worst Thing About My Sister*. Doubleday, 2012.

---. *This Girl*. Oxford University Press, 1988.

---. *Truth or Dare*. Macmillan, 1973.

---. *Twin Trouble*. Egmont, 1994.

---. *Vicky Angel*. Doubleday, 2000.

---. *Video Rose*. Puffin, 1992.

---. *Waiting for the Sky to Fall*. Oxford University Press, 1983.

---. *Wave Me Goodbye*. Doubleday, 2017.

---. *We Are the Beaker Girls*. Doubleday, 2019.

Woltz, Anna. *Aangespoeld*. Leopold, 2007.

---. *Alaska*. Querido, 2016.

---. *Alles kookt over*. Leopold, 2002.

---. *Black Box*. Leopold, 2007.

---. *De pizza spion*. Leopold, 2009.

---. *De tunnel*. Querido, 2021.

---. *Een hond met zakgeld*. Zwijsen, 2014.

---. *Evi, Nick en ik*. Leopold, 2011.

---. *Gips*. Querido, 2015.

---. *Haaientanden*. Stichting Collectieve Propaganda van het Nederlandse Boek, 2019.

---. *Het geheim van de stoere prinses*. Leopold, 2007.

---. *Het geheim van held nummer zes*. Leopold, 2011.

---. *Het geheim van ons vuur*. Leopold, 2004.

---. *Honderd uur nacht*. Querido, 2014.

---. *Ik kan nog steeds niet vliegen*. Leopold, 2012.

---. *Meisje nummer achttien*. Leopold, 2010.

---. *Meisje van Mars*. Querido, 2011.

---. *Mijn bijzonder rare week met Tess*. Querido, 2013.

---. *Onweer*. Leopold, 2009.

---. *Overleven in 4B*. Nijgh & Van Ditmar, 1998.

---. *Post uit de oorlog*. Leopold, 2006.

---. *Red mijn hond!*. Leopold, 2008.

---. *Tien dagen in een gestolen auto*. Leopold, 2008.

---. *Zondag Maandag Sterrendag*. Querido, 2017.

---. *Zusje te koop!*. Zwijsen, 2014.

Secondary works

Adolphs, Svenja. *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*. Routledge, 2006.

AFT. 'Schrijfster Hilde Vandermeeren Is Nieuwe Hoofdredacteur Kits'. *Het Nieuwsblad*, 21 Aug. 2007, p. 41.

Almond, David. *David Almond Interview*. Interview by Judith Ridge, Misrule, May 2003, http://misrule.com.au/wordpress/interviews/david-almond-interview/.

---. *Interview David Almond*. Interview by Living North, Dec. 2018.

---. *Interview with David Almond*. Interview by Vanessa Joosen, 13 Nov. 2020.

---. *January Interview*. Interview by Linda Richards, January Magazine, Feb. 2002, https://januarymagazine.com/profiles/almond.html.

---. 'Klaus Vogel and the Bad Lads'. *David Almond*, 10 Nov. 2017, www.davidalmond.com/books/klaus-vogel-and-the-bad-lads/.

Alvstad, Cecilia. 'Children's Literature and Translation'. *Handbook of Translation Studies*, edited by Yves Gambier and Luc Van Doorslaer, John Benjamins Publishing Company, 2010, pp. 22–27.

Anderson, Celia Catlett. *Style in Children's Literature: A Comparison of Passages From Books for Adults and for Children*. 1984. University of Rhode Island, PhD dissertation.

Anjirbag, Michelle Anya, and Frauke Pauwels. 'Constructing Age Transmedially: Framing Age in Text and on Screen in Skellig'. *Age in David Almond's Oeuvre: A Multi-Method Approach to Studying Age and the Life Course in Children's Literature*, Routledge, 2023, pp. 123–46.

Ansello, Edward F. 'Ageism: The Subtle Stereotype'. *Childhood Education*, vol. 54, no. 3, 1978, pp. 118–22, https://doi.org/10.1080/00094056.1978.10728378.

Apache. *OpenNLP*, 2022, https://opennlp.apache.org/.

Appleyard, J. A. *Becoming a Reader: The Experience of Fiction from Childhood to Adulthood*. Cambridge University Press, 1994.

Apseloff, Marilyn F. 'Grandparents in Fiction: A New Stereotype?' *Children's Literature Association Quarterly*, vol. 11, no. 2, 1986, pp. 80–82.

---. *They Wrote for Children Too: An Annotated Bibliography of Children's Literature by Famous Writers for Adults*. Greenwood Press, 1989.

Argamon, Shlomo, and Shlomo Levitan. *Measuring the Usefulness of Function Words for Authorship Attribution*, 2005.

Armstrong, Thomas. *The Human Odyssey: Navigating the Twelve Stages of Life*. Sterling, 2007.

Bailin, Alan, and Ann Grafstein. *Readability: Text and Context*. Palgrave Macmillan UK, 2016.

Bamman, David, Brendan O'Conner, and Noah A. Smith. 'Learning Latent Personas of Film Characters'. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2013, pp. 352–61, https://www.aclweb.org/anthology/P13-1035.

Bamman, David, Ted Underwood, and Noah A. Smith. 'A Bayesian Mixed Effects Model of Literary Character'. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 370–79.

Barnum, Phyllis Winet. 'Discrimination against the Aged in Young Children's Literature'. *The Elementary School Journal*, vol. 77, no. 4, 1977, pp. 301–06.

Beauvais, Clémentine. *The Mighty Child: Time and Power in Children's Literature*. John Benjamins Publishing Company, 2015.

Beckett, Sandra L. *Crossover Fiction: Global and Historical Perspectives*. Routledge, 2009.

---, editor. *Transcending Boundaries: Writing for a Dual Audience of Children and Adults*. Garland, 1999.

Behr, Kate. '"Same-as-Difference": Narrative Transformations and Intersecting Cultures in Harry Potter'. *Journal of Narrative Theory*, vol. 35, no. 1, 2005, pp. 112–32.

Beland, Robert M., and Terry L. Mills. 'Positive Portrayal of Grandparents in Current Children's Literature'. *Journal of Family Issues*, vol. 22, no. 5, 2001, pp. 639–51, https://doi.org/10.1177/019251301022005006.

Bennett, Jessica. 'I Am (an Older) Woman. Hear Me Roar'. *The New York Times*, 8 Jan. 2019.

Bertram, Bruce, Andee Rubin, and Kathleen Starr. 'Why Readability Formulas Fail'. *IEEE Transactions on Professional Communication*, vol. 1, 1981, pp. 50–52.

Best, Stephen, and Sharon Marcus. 'Surface Reading: An Introduction'. *Representations*, vol. 108, no. 1, 2009, pp. 1–21, https://doi.org/10.1525/rep.2009.108.1.1.

Binongo, José Nilo G. 'Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution'. *Chance*, vol. 16, no. 2, 2003, pp. 9–17.

Binongo, José Nilo G., and M. Wilfrid A. Smith. 'The Application of Principal Component Analysis to Stylometry'. *Literary and Linguistic Computing*, vol. 14, 1999, pp. 445–66.

Blake, Julie V. *What Did the National Curriculum Do for Poetry? Pattern, Prescription and Contestation in the Poetry Selected for GCSE English Literature 1988-2018*. 2019. University of Cambridge, PhD dissertation, https://doi.org/10.17863/CAM.47726.

Bland, Janice. *Children's Literature and Learner Empowerment: Children and Teenagers in English Language Education*. A&C Black, 2013.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 'Latent Dirichlet Allocation'. *Journal of Machine Learning Research*, vol. 3, 2003, pp. 993–1022.

Bode, Katherine. 'Why You Can't Model Away Bias'. *Modern Language Quarterly*, vol. 81, no. 1, 2020, pp. 95–124, https://doi.org/10.1215/00267929-7933102.

Bodner, Ehud. 'On the Origins of Ageism Among Older and Younger Adults'. *International Psychogeriatrics*, vol. 21, no. 6, 2009, pp. 1003–14, https://doi.org/10.1017/S104161020999055X.

Bottigheimer, Ruth B. 'An Important System of Its Own: Defining Children's Literature'. *The Princeton University Library Chronicle*, vol. 59, no. 2, 1998, pp. 191–210.

Bozkurt, Ilker Nadi, Özgür Bağhoğlu, and Erkan Uyar. 'Authorship Attribution'. *2007 22nd International Symposium on Computer and Information Sciences*, IEEE, 2007, pp. 1–5.

Brandwood, Leonard. ‘Stylometry and Chronology’. *The Cambridge Companion to Plato*, edited by Richard Kraut, Cambridge University Press, 1992, pp. 90–120.

Burnard, Lou. *What Is the Text Encoding Initiative*. OpenEdition Press, 2014.

Burrows, John. ‘All the Way Through: Testing for Authorship in Different Frequency Strata’. *Literary and Linguistic Computing*, vol. 22, no. 1, 2007, pp. 27–47.

---. *Computation into Criticism: A Study of Jane Austen’s Novels and an Experiment in Method*. Clarendon Press, 1987.

---. ‘“Delta”: A Measure of Stylistic Difference and a Guide to Likely Authorship’. *Literary and Linguistic Computing*, vol. 17, no. 3, 2002, pp. 267–87.

Butler, Rebecca. ‘The Literature Continuum: The Harry Potter Phenomenon’. *School Libraries Worldwide*, vol. 9, no. 1, 2003, pp. 64–77.

Bytheway, Bill. ‘Ageism’. *The Cambridge Handbook of Age and Ageing*, edited by Malcolm L. Johnson, Cambridge University Press, 2005, pp. 338–45.

Cadden, Mike. ‘Narrative Theory and Children’s Literature’. *A Companion to Children’s Literature*, edited by Karen Coats, Deborah Stevenson, and Vivian Yenika-Agbaw, John Wiley & Sons, 2022, pp. 376–89.

Čermáková, Anna. ‘Translating Children’s Literature: Some Insights from Corpus Stylistics’. *Ilha Do Desterro*, vol. 71, no. 1, 2018, pp. 117–34.

Čermáková, Anna, and Michaela Mahlberg. ‘Gender Inequality and Female Body Language in Children’s Literature’. *Digital Scholarship in the Humanities*, vol. 36, no. supplement_2, 2021, pp. ii72–ii77, https://doi.org/10.1093/llc/fqaa051.

Chai, Haixia, and Michael Strube. ‘Incorporating Centering Theory into Neural Coreference Resolution’. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2022, pp. 2996–3002, https://doi.org/10.18653/v1/2022.naacl-main.218.

Chaski, Carole E. ‘Empirical Evaluations of Language-Based Author Identification Techniques’. *Forensic Linguistics*, vol. 8, 2001, pp. 1–65.

Chivers, Sally. *From Old Woman to Older Women: Contemporary Culture and Women’s Narratives*. The Ohio State University Press, 2003.

Cinková, Silvie, and Jan Rybicki. ‘Stylometry in a Bilingual Setup’. *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, 2020, pp. 977–84.

Clark, Roger. ‘Why All the Counting? Feminist Social Science Research on Children’s Literature’. *Children’s Literature in Education*, vol. 33, 2002, pp. 285–95.

Cockrell, Amanda. ‘Harry Potter and the Secret Password: Finding Our Way in the Magical Genre’. *The Ivory Tower and Harry Potter: Perspectives in a Literary Phonomenon*, University of Missouri Press, 2002, pp. 15–26.

Coleman, Meri, and T. L. Liau. ‘A Computer Readability Formula Designed for Machine Scoring’. *Journal of Applied Psychology*, vol. 60, no. 2, 1975, pp. 283–84.

Collins, Michael. ‘Head-driven Statistical Models for Natural Language Parsing’. *Computational Linguistics*, vol. 29, no. 4, 2003, pp. 589–637.

Common Sense Media. ‘Harry Potter Age-by-Age Guide’. *Common Sense Media*, https://www.commonsensemedia.org/blog/harry-potter-age-by-age-guide. Accessed 28 July 2020.

---. ‘How We Rate and Review’. *Common Sense Media*, https://www.commonsensemedia.org/about- us/our-mission/about-our-ratings. Accessed 28 July 2020.

Covington, Michael A. *MATTR User Manual*. University of Georgia Artificial Intelligence Center, 2007.

Covington, Michael A., and Joe D. McFall. ‘Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)’. *Journal of Quantitative Linguistics*, vol. 17, no. 2, 2010, pp. 94–100, https://doi.org/10.1080/09296171003643098.

Crawford, Patricia A. ‘Crossing Boundaries: Addressing Ageism through Children’s Books’. *Horizons*, vol. 40, no. 3, 2000, pp. 161–74.

Cresci, Elena. ‘How Harry Potter Changed My Life’. *The Guardian: Opinion*, 17 Mar. 2016, https://www.theguardian.com/books/commentisfree/2016/mar/17/how-harry-potter-changed-my-life.

Crew, Hilary S. *Is It Really Mommie Dearest? Daughter-Mother Narratives in Young Adult Fiction*. Scarecrow Press, 2000.

Crisp, Thomas, Suzanne M. Knezek, Margaret Quinn, Gary E. Bingham, Kristy Girardeau, Francheska Starks. ‘What’s on Our Bookshelves? The Diversity of Children’s Literature in Early Childhood Classroom Libraries’. *Journal of Children’s Literature*, vol. 42, no. 2, 2016, p. 29.

Cross, Amy, Cherie Allan, and Kerry Kilner. ‘Digital Curation, AustLit, and Australian Children’s Literature’. *International Research in Children’s Literature*, vol. 12, no. 1, 2019, pp. 1–17.

Crossley, Scott A., Stephen Skalicky, and Mihai Dascalu. ‘Moving Beyond Classic Readability Formulas: New Methods and New Models’. *Journal of Research in Reading*, vol. 0, no. 0, 2019, pp. 1–21.

Cummings, Elaine, and William Henry. *Growing Old: The Process of Withdrawal*. Basic Books, 1961.

Cunningham, Diana M. H., and Kalina Bontcheva. *Text Processing with GATE (Version 6)*. University of Sheffield D, 2011.

Da, Nan Z. ‘The Computational Case against Computational Literary Studies’. *Critical Inquiry*, vol. 45, no. 3, 2019, pp. 601–39, https://doi.org/10.1086/702594.

Dale, Edgar, and Jeanne S. Chall. ‘A Formula for Predicting Readability: Instructions’. *Educational Research Bulletin*, 1948, pp. 37–54.

---. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, 1995.

De Langhe, Loic, Thierry Desot, Orphée De Clerq, and Veronique Hoste. ‘A Benchmark for Dutch End-to-End Cross-Document Event Coreference Resolution’. *Electronics*, vol. 12, no. 4, 2023, p. 850, https://doi.org/10.3390/electronics12040850.

De Vriend, Gerard. 'Literaire Kritiek of Consumentenvoorlichting?' *Literatuur Zonder Leeftijd*, vol. 20, 2006, pp. 24–34.

Diederich, Joachim, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 'Authorship Attribution with Support Vector Machines'. *Applied Intelligence*, vol. 19, 2003, pp. 109–23.

Diekman, Amanda B., and Sarah K. Murnen. 'Learning to Be Little Women and Little Men: The Inequitable Gender Equality of Nonsexist Children's Literature'. *Sex Roles*, vol. 50, 2004, pp. 373–85, https://doi.org/10.1023/B:SERS.0000018892.26527.e.

Drucker, Johanna. 'Why Distant Reading Isn't'. *Publications of the Modern Language Association of America*, vol. 132, no. 3, 2017, pp. 628–35, https://doi.org/10.1632/pmla.2017.132.3.628.

Duffy, Thomas M. 'Readability Formulas: What's the Use?' *Designing Usable Texts*, edited by Thomas M. Duffy and Robert Waller, Academic Press, 1985, pp. 113–43.

Duthoy, Leander. 'An Exploration of Reader-Response Research Through My Name Is Mina'. *Age in David Almond's Oeuvre: A Multi-Method Approach to Studying Age and the Life Course in Children's Literature*, Routledge, 2023, pp. 92–122.

Eccleshare, Julia. *A Guide to the Harry Potter Novels*. Continuum, 2002.

Eco, Umberto. *The Role of the Reader: Explorations in the Semiotics of Texts*. Indiana University Press, 1979.

Eder, Maciej. 'Does Size Matter? Authorship Attribution, Small Samples, Big Problem'. *Digital Scholarship in the Humanities*, vol. 30, no. 2, 2015, pp. 167–82.

Eder, Jens, Fotis Jannidis, and Ralf Schneider, editors. *Characters in Fictional Worlds: Understanding Imaginary Beings in Literature, Film, and Other Media*. De Gruyter, 2010.

Errington, Philip W. *J.K. Rowling: A Bibliography 1997-2013*. Bloomsbury Academic, 2015.

Eve, Paul Martin. *Close Reading with Computers: Textual Scholarship, Computational Formalism, and David Mitchell's Cloud Atlas*. Stanford University Press, 2019, https://doi.org/10.21627/9781503609372.

Ewers, Hans-Heino. *Fundamental Concepts of Children's Literature Research: Literary and Sociological Approaches*. Routledge, 2009.

Fastenau, Kyra. 'Listen En Lagen: De Volwassenwording van de Jeugdliteratuur'. *De Leeswelp*, 2014, p. 36.

Fernández-Ballesteros, Rocío, Jean Marie Robine, Alan Walker, and Alex Kalache. 'Active Aging: A Global Goal'. *Current Gerontology and Geriatrics Research*, 2013, pp. 1–4.

Fine, Anne. *Interview with Anne Fine*. Interview by Vanessa Joosen, 27 Oct. 2020.

Fitzsimmons, Rebekah, and Suzan Alteri. "Possibly Impossible; Or, Teaching Undergraduates to Confront Digital and Archival Research Methodologies, Social Media Networking, and Potential Failure". *Journal of Interactive Technology and Pedagogy*, vol. 14, 2019.

Flamion, Allison, Pierre Missotten, Lucie Jennotte, Noémie Hody, and Stéphane Adam. 'Old Age-Related Stereotypes of Preschool Children'. *Frontiers in Psychology*, vol. 11, no. 807, 2020, pp. 1–13, https://doi.org/10.3389/fpsyg.2020.00807.

Flekova, Lucie, and Iryna Gurevych. 'Personality Profiling of Fictional Characters Using Sense-Level Links between Lexical Resources'. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2015, pp. 1805–15.

Forsyth, Richard S., David I. Holmes, and E. K. Tse. 'Cicero, Sigonio, and Burrows: Investigating the Authenticity of the Consolatio'. *Literary and Linguistic Computing*, vol. 14, no. 3, 1999, pp. 375–400.

Fraboni, Maryann, Robert Saltstone, and Susan Hughes. 'The Fraboni Scale of Ageism (FSA): An Attempt at a More Precise Measure of Ageism'. *Canadian Journal on Aging*, vol. 9, no. 1, 1990, pp. 56–66.

Fraser, Lindsey. 'Review: Harry Potter and the Philosopher's Stone'. *The Scotsman*, 28 June 1997.

Fry, Edward. 'Readability versus Leveling'. *The Reading Teacher*, vol. 56, no. 3, 2002, pp. 286–91.

Galef, David. 'Crossing Over: Authors Who Write Both Children's and Adult's Fiction'. *Children's Literature Association Quarterly*, vol. 20, no. 1, 1995, pp. 29–35. https://doi.org/10.1353/chq.0.0905.

Gavin, Adrienne, editor. *The Child in British Literature: Literary Constructions of Childhood, Medieval to Contemporary*. Palgrave Macmillan, 2012.

Ghesquière, Rita. *Het Verschijnsel Jeugdliteratuur*. Acco, 2000.

Ghesquière, Rita, Vanessa Joosen, and Helma van Lierop. *Een Land van Waan En Wijs. Geschiedenis van de Nederlandse Jeugdliteratuur*. Atlas Contact, 2014.

Giddens, Eugene. "Distant Reading and Children's Literature." *The Edinburgh Companion to Children's Literature*, edited by Clémentine Beauvais and Maria Nikolajeva. Edinburgh University Press, 2017, pp. 305–13.

Green, Lorraine. *Understanding the Life Course: Sociological and Psychological Perspectives*. Polity, 2010.

Gregori-Signes, Carmen, and Begoña Clavel-Arroitia. 'Analysing Lexical Density and Lexical Diversity in University Students' Written Discourse'. *Procedia-Social and Behavioral Sciences*, vol. 198, 2015, pp. 546–56.

Grieve, Jack. 'Quantitative Authorship Attribution: An Evaluation of Techniques'. *Literary and Linguistic Computing*, vol. 22, no. 3, 2007, pp. 251–70.

Griffin, Melanie. *Once Upon a Genre: Distant Reading, the Newbery Medal, and the Affordances of Interdisciplinary Paradigms for Understanding Children's Literature*. 2018. University of South Florida, PhD dissertation.

Grit, Diederik. 'Henk van Kerkwijk En Zijn Lezers'. *Project Jeugdliteratuur*, 2.0.02, 1977.

Gullette, Margaret Morganroth. *Aged by Culture*. University of Chicago Press, 2004.

---. *Agewise: Fighting the New Ageism in America*. University of Chicago Press, 2011.

---. *Ending Ageism or How Not to Shoot Old People*. Rutgers University Press, 2017.

Gunning, Robert. *The Technique of Clear Writing*. McGraw-Hill, 1952.

Guttery, Jean. 'Style in Children's Literature'. *The Elementary English Review*, vol. 18, no. 6, 1941, pp. 208–12.

Hanson, Michelle. 'Older Women: Has Society Forgotten How to Value Them?' *The Guardian*, 22 Apr. 2013, https://www.theguardian.com/lifeandstyle/2013/apr/22/older-women-society-value-them.

Haverals, Wouter, Lindsey Geybels, and Vanessa Joosen. 'A Style for Every Age: A Stylometric Enquiry into Crosswriters for Children, Adolescents and Adults'. *Language and Literature*, vol. 31, no. 1, 2022, pp. 62–84.

Haverals, Wouter, and Lindsey Geybels. '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'. *Journal of Cultural Analytics*, vol. 6, no. 1, 2021, pp. 255–84.

Haverals, Wouter, and Vanessa Joosen. 'Constructing Age in Children's Literature: A Digital Approach to Guus Kuijer's Oeuvre'. *The Lion and the Unicorn*, vol. 45, 2021, pp. 25–45, https://doi.org/10.1353/uni.2021.0002.

Hearn, Jeff. 'Imaging the Aging of Men'. *Images of Aging: Cultural Representations of Later Life*, edited by Mike Featherstone and Andrew Wernick, Routledge, 1995, pp. 97–114.

Henneberg, Sylvia. 'Moms Do Badly, But Grandmas Do Worse: The Nexus of Sexism and Ageism in Children's Classics'. *Journal of Aging Studies*, vol. 24, 2010, pp. 125–34, https://doi.org/10.1016/j.jaging.2008.10.003.

---. 'Of Creative Crones and Poetry: Developing Age Studies Through Literature'. *NWSA Journal*, vol. 18, no. 1, 2006, pp. 106–25.

Herman, Luc, and Bart Vervaeck. *Handbook of Narrative Analysis*. University of Nebraska Press, 2005.

---.'Ideology and Narrative Fiction'. *The Living Handbook of Narratology*, edited by Peter Hühn et al, 2013.

Herrmann, J. Berenike, Christof Schöch, and Karina van Dalen-Oskam. 'Revisiting Style, a Key Concept in Literary Studies'. *Journal of Literary Theory*, vol. 9, no. 1, 2015, https://doi.org/10.1515/jlt-2015-0003.

Herrmann, J. Berenike, Anne-Sophie Bories, Francesca Frontini, Simone Rebora, and Jan Rybicki. 'Response by the Special Interest Group on Digital Literary Stylistics to Nan Z. Da's Study'. *Journal of Cultural Analytics*, vol. 5, no. 1, 2020, https://doi.org/10.22148/001c.11827.

Hillman, Judith S. 'An Analysis of Male and Female Roles in Two Periods of Children's Literature'. *Journal of Educational Research*, vol. 68, no. 2, 1974, pp. 84–88, https://doi.org/10.1080/00220671.1974.10884713.

Hockey, Jenny, and Allison James. 'Back to Our Futures: Imaging Second Childhood'. *Images of Aging: Cultural Reflections of Later Life*, edited by Mike Featherstone and Andrew Wernick, Routledge, 1995, pp. 135–48.

Hollindale, Peter. *Ideology and the Children's Book*. Thimble Press, 1988.

Hollindale, Peter. *Signs of Childness in Children's Books*. Thimble Press, 2001.

Holmes, David I. 'The Analysis of Literary Style - A Review'. *Journal of the Royal Statistical Society*, vol. 148, no. 4, 1985, pp. 328–41.

---. 'The Evolution of Stylometry in Humanities Scholarship'. *Literary and Linguistic Computing*, vol. 13, no. 3, 1998, pp. 111–17.

Honnibal, Matthew, and Ines Montani. *SpaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing*, 2017.

Hoover, David L. 'Multivariate Analysis and the Study of Style Variation'. *Literary and Linguistic Computing*, vol. 18, no. 4, 2003, pp. 341–60, https://doi.org/10.1093/llc/18.4.341.

---. 'Statistical Stylistics and Authorship Attribution: An Empirical Investigation'. *Literary and Linguistic Computing*, vol. 16, no. 4, 2001, pp. 421–44.

Hunt, Peter. 'What the Authors Tell Us'. *International Companion Encyclopedia of Children's Literature*, edited by Peter Hunt, Routledge, 1996, pp. 548–63.

Hurkmans, Nicole. 'Waar Liggen de Grenzen Nog? Een Onderzoek Naar Vier Dubbelpublieksauteurs'. *Literatuur Zonder Leeftijd*, vol. 22, 2008, pp. 98–114.

Immel, Andrea. 'Children's Books and Constructions of Childhood'. *The Cambridge Companion to Children's Literature*, edited by M.O. Grenby and Andrea Immel, Cambridge University Press, 2009, pp. 19–34.

Iser, Wolfgang. *The Implied Reader: Patterns of Communication in Prose Fiction from Bunyan to Beckett*. Johns Hopkins University Press, [1972] 1974.

Jacobs, Arthur M. 'Sentiment Analysis for Words and Fiction Characters from the Perspective of Computational (Neuro-)Poetics'. *Frontiers in Robotics and AI*, vol. 6, 2019, pp. 1–13, https://doi.org/10.3389/frobt.2019.00053.

Janelli, Linda M. 'Depictions of Grandparents in Children's Literature'. *Educational Gerontology*, vol. 14, no. 3, 1988, pp. 193–202.

Jänicke, Stefan, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 'On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges'. *EuroVis*, 2015, pp. 83–103.

Jannidis, Fotis. 'Character'. *The Living Handbook of Narratology*, edited by Peter Hühn et al, 2013.

---. 'On the Perceived Complexity of Literature. A Response to Nan Z. Da'. *Journal of Cultural Analytics*, vol. 5, no. 1, 2020, https://doi.org/10.22148/001c.11829.

Jannidis, Fotis, Steffen Pielström, Christof Schöch, and Thorsten Vitt. *Improving Burrow's Delta: An Empirical Evaluation of Text Distance Measures*. 2015, pp. 1–10.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

Johansson, Victoria. 'Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective'. *Working Papers*, vol. 53, 2008, pp. 61–79.

Johnson, Wendell. *Language and Speech Hygiene*, 1939.

Joosen, Vanessa. *Adulthood in Children's Literature*. Bloomsbury Academic, 2018.

---. 'Aging in Children's Literature'. *Encyclopedia of Gerontology and Population Aging*, edited by Danan Gu and Matthew E. Dupre, Springer International, 2020, pp. 1–5, https://doi.org/10.1007/978-3-319-69892-2_250-1.

---. '*Counting Stars*, Discounting Years? Life Writing and Memory Studies'. *Age in David Almond's Oeuvre: A Multi-Method Approach to Studying Age and the Life Course in Children's Literature*, Routledge, 2023, pp. 12–35.

---. ‘Grandparents’. *Routledge Companion to Family*, Routledge, forthcoming.

---. *Hoe Oud Is Jong? Leeftijd En Jeugdliteratuur*. Letterwerk, 2022.

---. ‘Not Your (Ordinary) Grandma: Old Age in Three Contemporary Dutch Children’s Books’. *Children’s Literature and Intergenerational Relationships: Encounters of the Playful Kind*, edited by Justyna Deszcz-Tryhubczak and Irena Barbara Kalla, Palgrave Macmillan, 2021, pp. 181–95.

---. ‘Research in Action: Constructing Age for Young Readers’. *International Research in Children’s Literature*, vol. 14, no. 3, 2021, pp. 252–68.

---. ‘Second Childhoods and Intergenerational Dialogues: How Children’s Literature Studies and Age Studies Can Supplement Each Other’. *Children’s Literature Association Quarterly*, vol. 40, no. 2, 2015, pp. 126–40.

---. ‘Writing When Young: Bart Moeyaert as a Young Adult Author’. *European Journal of Life Writing*, vol. 10, 2021, pp. BB65–83, https://doi.org/10.21827/ejlw.10.38163.

Juola, Patrick. ‘How a Computer Program Helped Show J.K. Rowling Write A Cuckoo’s Calling’. *Scientific American*, 20 Aug. 2013, https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/?print=true.

Karsdorp, Folgert. *Retelling Stories: A Computational-Evolutionary Perspective*. 2016. Radboud Universiteit, PhD dissertation.

Karsdorp, Folgert, and Antal van den Bosch. ‘Identifying Motifs in Folktales Using Topic Models’. *Proceedings of BENELEARN*, 2013.

Karsdorp, Folgert, Enrique Manjavacas, and Lauren Fonteyn. ‘Introducing Functional Diversity: A Novel Approach to Lexical Diversity in (Historical) Corpora’. *Proceedings of the Computational Humanities Research Conference 2022*, CEUR Workshop Proceedings, 2022, pp. 114–26.

Kessler, Jason S. ‘Scattertext: A Browser-Based Tool for Visualizing How Corpora Differ’. *ArXiv*, vol. 1703.00565 [Cs], 2017, https://arxiv.org/pdf/1703.00565.pdf.

Kestemont, Mike. ‘Function Words in Authorship Attribution: From Black Magic to Theory?’ *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, Association for Computational Linguistics, 2014, pp. 59–66.

Kestemont, Mike, and Luc Herman. ‘Can Machines Read (Literature)?’ *Umanistica Digitale*, vol. 5, 2019, https://doi.org/10.6092/ISSN.2532-8816/8511.

Kim, Evgeny, and Roman Klinger. ‘Frowning Frodo, Wincing Leia, and a Seriously Great Friendship: Learning to Classify Emotional Relationships of Fictional Characters’. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 647–53, https://doi.org/10.48550/arXiv.1903.12453.

Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. ‘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.

Kitaev, Nikita, and Dan Klein. ‘Constituency Parsing with a Self-Attentive Encoder’. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Vol 1*, 2018.

Kjetssa, Geir. ‘And Quiet Flows the Don Through the Computer’. *Association for Literary and Linguistic Computing Bulletin*, vol. 7, 1979, pp. 248–56.

Knowles, Murray, and Kirsten Malmkjaer. Language and Control in Children's Literature. Routledge, 1996.

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*, vol. 33, no. 1, 2018, pp. 59–71.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 'Automatically Categorizing Written Texts by Author Gender'. *Literary and Linguistic Computing*, vol. 17, no. 4, 2002, pp. 401–12, https://doi.org/10.1093/llc/17.4.401.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 'Computational Methods in Authorship Attribution'. *Journal of the American Society for Information Science and Technology*, vol. 60, 2009, pp. 9–26.

Kortemark, R. 'Vandermeeren Debuteert Met Jeugdboek'. *Het Nieuwsblad*, 1 June 2001, p. 15.

Krashen, Stephen. 'The Lexile Framework: Unnecessary and Potentially Harmful'. *CSLA Journal*, vol. 24, no. 2, 2001, pp. 25–26.

Kuijer, Guus. *Het Geminachte Kind*. Synopsis, 1980.

Kümmerling-Meibauer, Bettina. 'Seriality in Children's Literature'. *The Edinburgh Companion to Children's Literature*, edited by Clémentine Beauvais and Maria Nikolajeva, Edinburgh University Press, 2017, pp. 167–78.

Kyle, Kristopher. 'Measuring Lexical Richness'. *The Routledge Handbook of Vocabulary Studies*, edited by Stuart Webb, Routledge, 2019, pp. 454–76.

Lacoss, Jann. 'Of Magicals and Muggles: Reversals and Revulsions at Hogwarts'. *The Ivory Tower and Harry Potter: Perspectives in a Literary Phenomenon*, edited by Lana Whited, University of Missouri Press, 2002, pp. 67–88.

Łajewska, Weronika, and Anna Wróblewska. 'Protagonists' Tagger in Literary Domain—New Datasets and a Method for Person Entity Linkage'. *ArXiv*, vol. 2110.01349, 2021.

Lane, Melissa. 'Socrates and Plato: An Introduction'. *The Cambridge History of Greek and Roman Political Thought*, edited by Christopher J. Rowe, Malcolm Schofield, Simon Harrison, and Melissa Lane, Cambridge University Press, 2000, pp. 155–63.

Latham, Don. *David Almond: Memory and Magic*. Scarecrow Press, 2006.

Lem, Ellyn A. *Gray Matters: Finding Meaning in the Stories of Later Life*. Rutgers University Press, 2020.

Lennon, Colleen, and Hal Burdick. *The Lexile Framework as an Approach for Reading Measurement and Success*, 2004.

Lerer, Seth. *Children's Literature: A Reader's History, from Aesop to Harry Potter*. University of Chicago Press, 2009.

Levitsky, Viktor V., and Yulia P. Melnyk. 'Sentence Length and Sentence Structure in English Prose'. *Glottometrics*, vol. 21, 2011, pp. 14–24.

Levy, Becca. 'Stereotype Embodiment: A Psychosocial Approach to Aging'. *Current Directions in Psychological Science*, vol. 18, no. 6, 2009, pp. 332–36, https://doi.org/10.1111/j.1467-8721.2009.01662.x.

Lister, D. ‘Pullman Wins Whitbread for Children’s Fantasy’. *Independent*, 23 Jan. 2002.

Loper, Edward, and Steven Bird. ‘NLTK: The Natural Language Toolkit’. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, 2002, pp. 63–70, https://doi.org/10.3115/1118108.1118117.

López-Escobedo, Fernanda, Carlos-Francisco Méndez-Cruz, Gerardo Sierra, and Julián Solórzano-Soto. ‘Analysis of Stylometric Variables in Long and Short Texts’. *Procedia-Social and Behavioral Sciences*, vol. 95, 2013, pp. 604–11.

Mahlberg, Michaela, Peter Stockwell, Johan de Joode, Catherine Smith, and Matthew Brook O’Donnell. ‘CliC Dickens: Novel Uses of Concordances for the Integration of Corpus Stylistics and Cognitive Poetics’. *Corpora*, vol. 11, no. 3, 2016, pp. 433–63.

Mahlberg, Michaela, and Viola Wiegand. ‘Literary Stylistics’. *The Routledge Handbook of English Language and Digital Humanities*, edited by Svenja Adolphs and Dawn Knight, Routledge, 2020, pp. 306–27.

Mailloux, Steven J. *Interpretive Conventions: The Reader in the Study of American Fiction*. Cornell University Press, 1982.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. ‘The Stanford CoreNLP Natural Language Processing Toolkit’. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 55–60.

Mannion, David, and Peter Dixon. ‘Sentence-Length and Authorship Attribution: The Case of Oliver Goldsmith’. *Literary and Linguistic Computing*, vol. 19, no. 4, 2004, pp. 497–508, https://doi.org/10.1093/llc/19.4.497.

Martindale, Colin, and Dean McKenzie. ‘On the Utility of Content Analysis in Author Attribution: The Federalist’. *Computers and the Humanities*, vol. 29, no. 4, 1995, pp. 259–70.

McCarthy, Philip, and Scott Jarvis. ‘Vocd: A Theoretical and Empirical Evaluation’. *Language Testing*, vol. 24, no. 4, 2007, pp. 459–88.

McClosky, David, Eugene Charniak, and Mark Johnson. ‘Effective Self-training for Parsing’. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006, pp. 152–159.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajic. ‘Non-projective Dependency Parsing Using Spanning Tree Algorithms’. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 523–530.

McGuire, Sandra L. ‘Early Children’s Literature and Aging’. *Creative Education*, vol. 7, 2016, pp. 2604–12, https://doi.org/10.4236/ce.2016.717245.

McIntyre, Dan, and Brian Walker. *Corpus Stylistics: Theory and Practice*. Edinburgh University Press, 2019.

McLaughlin, G. Harry. ‘SMOG Grading: A New Readability Formula’. *Journal of Reading*, vol. 12, no. 8, 1969, pp. 639–46.

Mendenhall, Thomas C. ‘The Characteristic Curves of Composition’. *Science*, vol. ns-9, no. 214s, 1887, pp. 237–46, https://doi.org/10.1126/science.ns-9.214S.237.

Mesmer, Heidi Anne E. *Tools for Matching Readers to Texts. Research-Based Practices*. Guilford Press, 2008.

Meyer, Bonnie J. F. 'Identification of the Structure of Prose and Its Implications for the Study of Reading and Memory'. *Journal of Reading Behavior*, vol. 7, no. 1, 1975, pp. 7–47.

Molenberghs, Geert, Luc Bijnens, and David Shaw. 'Linear Mixed Models and Missing Data'. *Linear Mixed Models in Practice*, Springer, 1997, pp. 191–274.

Montepare, Joann M., and Leslie A. Zebrowitz. 'A Social-Developmental View of Ageism'. *Ageism: Stereotyping and Prejudice against Older Persons*, edited by Todd D. Nelson, MIT Press, 2002, pp. 77–125.

Moretti, Franco. *Distant Reading*. Verso, 2013.

Morrow, Virginia. 'What's in a Number? Unsettling the Boundaries of Age'. *Childhood*, vol. 20, no. 2, 2013, pp. 151–55.

Mosteller, Frederick, and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison Welsey, 1964.

My Byline Media. 'Readability Formulas'. *Readability Formulas*, https://readabilityformulas.com/search/pages/Readability_Formulas/. Accessed 24 Aug. 2020.

Nalisnick, Eric T., and Henry S. Baird. 'Character-to-Character Sentiment Analysis in Shakespeare's Plays'. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. Volume 2 (Short Papers), Association for Computational Linguistics, 2013, pp. 479–83.

Nel, Philip. *J.K. Rowling's Harry Potter Novels: A Reader's Guide*. A&C Black, 2001.

Ng, Reuben, Heather G. Allore, Mark Trentalange, Joan K. Monin, and Becca R. Levy. 'Increasing Negativity of Age Stereotypes across 200 Years: Evidence from a Database of 400 Million Words'. *PLoS ONE*, vol. 10, no. 2, 2015, https://doi.org/10.1371/journal.pone.0117086.

Nikolajeva, Maria. 'The Changing Aesthetics of Character in Children's Fiction'. *Style*, vol. 35, no. 3, 2001, pp. 430–53.

---. *Children's Literature Comes of Age*. Garland Publishing, 1996.

---. 'Narrative Theory and Children's Literature'. *International Companion Encyclopedia of Children's Literature*, edited by Peter Hunt, Routledge, 2004, pp. 166–78.

---. *Power, Voice and Subjectivity in Literature for Young Readers*. Routledge, 2010.

---. *The Rhetoric of Character in Children's Literature*. Scarecrow Press, 2002.

---. 'What Is It Like to Be a Child? Childness in the Age of Neuroscience'. *Children's Literature in Education*, vol. 50, no. 1, 2019, pp. 23–37.

Nikolajeva, Maria, and Carole Scott. *How Picturebooks Work*. Garland, 2001.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 'Maltparser: A Language-Independent System for Data-Driven Dependency Parsing'. *Natural Language Engineering*, vol. 13, no. 2, 2007, pp. 95–135.

Nodelman, Perry. *The Hidden Adult: Defining Children's Literature*. John Hopkins University Press, 2008.

Noecker Jr, John, Michael Ryan, and Patrick Juola. 'Psychological Profiling Through Textual Analysis'. *Literary and Linguistic Computing*, vol. 28, no. 3, 2013, pp. 382–87.

Oelke, Daniela, Dimitrios Kokkinakis, and Daniel A. Keim. 'Fingerprint Matrices: Uncovering the Dynamics of Social Networks in Prose Literature'. *Computer Graphics Forum*, vol. 32, no. 3, 2013, pp. 371–80, https://doi.org/10/1111/cgf.12124.

Park, SunYoung, Kyuri Choi, Haeun Yu, and Youngjoong Ko. 'Never Too Late to Learn: Regularizing Gender Bias in Coreference Resolution'. *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, ACM, 2013, pp. 15–23, https://doi.org/10.1145/3539597.3570473.

Pennebaker, James W. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press, 2011. https://doi.org/10.1016/S0262-4079(11)62167-2.

Petrov, Slav, Leon Barrett, Romain Thibaux, and Dan Klein. 'Learning Accurate, Compact, and Interpretable Tree Annotation'. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 433–40.

Pickard, Susan. *Age Studies: A Sociological Examination of How We Age and Are Aged Through the Life Course*. Sage, 2016.

Pifer, Ellen. 'The Image of Childhood, Present and Past'. *Demon or Doll: Images of the Child in Contemporary Writing and Culture*, University Press of Virginia, 2000, pp. 11–27.

Poix, Céline. 'Neology in Children's Literature: A Typology of Occasionalisms'. *Lexis*, no. 12, 2018.

Potter, Rosanne G. 'Statistical Analysis of Literature: A Retrospective on Computers and the Humanities, 1966-1990'. *Computers and the Humanities*, vol. 25, no. 6, 1991, pp. 401–29.

Prince, Gerald. 'Reader'. *Handbook of Narratology*, edited by Peter Hühn, John Pier, Wolf Schmid, and Jörg Schönert, Walter de Gruyter, 2009, pp. 398–410.

Pullman, Philip. *No to Age Banding: Philip Pullman's Address*. Society of Authors' Conference, Cambridge, 31 Aug. 2008.

---. 'Questions & Answers'. *Philip Pullman*, http://www.philip-pullman.com/QAs. Accessed 11 Sept. 2020.

Radl, Jonas. 'Too Old to Work, or Too Young to Retire? The Pervasiveness of Age Norms in Western Europe'. *Work, Employment and Society*, vol. 26, no. 5, 2012, pp. 755–71.

Rassuli, Kathleen M. 'Evidence of Marketing Strategy in the Early Printed Book Trade: An Application of Hollander's Historical Approach'. *Historical Perspectives in Marketing: Essays in Honor of Stanley C. Hollander*, edited by Terence Nevett and Ronald A. Fullerton, D.C. Heath, 1988, pp. 91–107.

Rimmon-Kenan, Shlomith. *Narrative Fiction: Contemporary Poetics*. 2nd ed., Routledge, 2002.

Rochat, Yannick, and Frédéric Kaplan. 'Analyse Des Réseaux de Personnages Dans Les Confessions de Jean-Jacques Rousseau'. *Les Cahiers Du Numérique*, vol. 10, no. 3, 2014, pp. 109–33, https://doi.org/10.3166/LCN.10.3.109-133.

Rose, Jacqueline. *The Case of Peter Pan or the Impossibility of Children's Fiction*. Macmillan, 1994.

Rowe, Rebecca, Tolonda Henderson, and Tianyu Wang. 'Text Mining, Hermione Granger, and Fan Fiction: What's in a Name?' *Transformative Works and Cultures*, vol. 36, 2021, https://doi.org/10.3983/twc.2021.1997.

Rowling, J. K. *J.K. Rowling Introduces the Ickabog*. 26 May 2020, https://www.jkrowling.com/j-k-rowling-introduces-the-ickabog/.

Rudvin, Mette. 'Translation and "Myth": Norwegian Children's Literature in English'. *Perspectives*, vol. 2, no. 2, 1994, pp. 199–211.

Sabbe, J. 'Hilde Vandermeeren Schrijft Thriller Voor Volwassen Beginnende Lezers'. *Krant van West-Vlaanderen*, 12 July 2019, p. 10.

Sánchez-Eppler, Karen. 'Childhood'. *Keywords for Children's Literature*, edited by Lissa Paul and Philip Nel, New York University Press, 2011, pp. 35–41.

Sargeant, Betty. *How Far is Up? The Functional Properties and Aesthetic Materiality of Children's Storybook Applications*. 2015. RMIT University, PhD dissertation.

Sax, Aline. *Interview with Aline Sax*. Interview by Vanessa Joosen, 11 Feb. 2021.

Schmid, Wolf. 'Implied Reader'. *The Living Handbook of Narratology*, edited by Peter Hühn et al, 2014.

Schöch, Christof, Karina van Dalen-Oskam, Maria Antoniak, Fotis Jannidis, and David Mimno. 'Replication and Computational Literary Studies'. *Digital Humanities Conference 2020*, Ottowa, Canada, June 2020, https://doi.org/10.5281/ZENODO.3893428.

Scholastic. 'The Ultimate Harry Potter Book List'. *Scholastic Teaching Tools*, 2020, https://www.scholastic.com/teachers/teaching-tools/book-lists/the-complete-harry-potter-book-list.html.

Schriver, Karen A. 'Readability Formulas in the New Millennium: What's the Use?' *ACM Journal of Computer Documentation*, vol. 24, no. 3, 2000, pp. 138–40.

Segal, Lynne. 'The Coming of Age Studies'. *Age, Culture, Humanities: An Interdisciplinary Journal*, vol. 1, 2014, pp. 31–34. https://doi.org/10.7146/ageculturehumanities.v1i.129944.

Sergeant, Betty. *How Far is Up? The Functional Properties and Aesthetic Materiality of Children's Storybook Applications*. RMIT University, 2015.

Shavit, Zohar. 'The Double Attribution of Texts for Children and How It Affects Writing for Children'. *Transcending Boundaries: Writing for a Dual Audience of Children and Adults*, edited by Sandra L. Beckett, Garland, 1999, pp. 83–98.

---. *Poetics of Children's Literature*. University of Georgia Press, 2009.

Sikorska, Karolina. *All Children, Except One, Grow Up: A Stylometric Analysis of Polish Translations of JM Barrie's Peter Pan Stories*. 2013. Jagiellonian University, PhD dissertation.

Silva, Emma-Louise. 'Social and Material Minds Through the Lens of Cognitive Narratology in Clay and Bone Music'. *Age in David Almond's Oeuvre: A Multi-Method Approach to Studying Age and the Life Course in Children's Literature*, Routledge, 2023, pp. 36–58.

Smeets, Roel. *Character Constellations: Representations of Social Groups in Present-Day Dutch Literary Fiction*. 2020. Leuven University, PhD dissertation.

Smith, Edgar A., and R. J. Senter. 'Automated Readability Index'. *Aerospace Medical Research Laboratories*, 1967, pp. 1–14.

Smith, Peter W. H., and W. Aldridge. 'Improving Authorship Attribution. Optimizing Burrows' Delta Method'. *Journal of Quantitative Linguistics*, vol. 18, no. 1, 2011, pp. 63–88, https://doi.org/10.1080/09296174.2011.533591.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 'Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank'. *EMNLP*, 2013.

Stamatatos, Efstathios. 'A Survey of Modern Authorship Attribution Methods'. *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, 2009, pp. 538–56.

Stephens, John. *Language and Ideology in Children's Fiction*. Longman, 1992.

---. 'Linguistics and Stylistics'. *International Companion Encyclopedia of Children's Literature*, edited by Peter Hunt, Routledge, 2004, pp. 99–111.

---. 'Maintaining Distinctions: Realism, Voice, and Subject Position in Australian Young Adult Fiction'. *Transcending Boundaries: Writing for a Dual Audience of Children and Adults*, edited by Sandra L. Beckett, Garland, 1999, pp. 183–98.

Stoffel, Florian, Wolfgang Jentner, Michael Behrisch, Johannes Fuchs, and Daniel Keim. 'Interactive Ambiguity Resolution of Named Entities in Fictional Literature'. *Computer Graphics Forum*, vol. 36, no. 3, 2017, pp. 189–200, https://doi.org/10.1111/cgf.13179.

Tallentire, D. Roger. *An Appraisal of Methods and Models in Computational Stylistics, with Particular Reference to Author Attribution*. 1972. University of Cambridge, PhD dissertation.

---. 'Confirming Intuitions About Style, Using Concordances'. *The Computer in Literary and Linguistic Studies*, edited by A. Jones and R. F. Churchhouse, University of Wales Press, 1976, pp. 309–38.

Talley, Lee A. 'Young Adult'. *Keywords for Children's Literature*, edited by Philip Nel and Lissa Paul, New York University Press, 2011, pp. 228–32.

Tedeschi, Simone, and Roberto Navigli. *MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation)*. Unpublished, 2022.

TEI Consortium, editor. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 4.4.0, 2022.

Temple, J. T. 'A Multivariate Synthesis of Published Platonic Stylometric Data'. *Literary and Linguistic Computing*, vol. 11, no. 2, 1996, pp. 67–75.

Templin, Mildred C. *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press, 1957.

The Page Turner. 'Author Sets Up Book Cover Protest'. *Oxford Mail*, 10 June 2008.

Thomas, Ed. 'Word Frequency and Collocation: Using Children's Literature in Adult Learning'. *ICAME Journal*, vol. 39, no. 1, 2015, pp. 85–110, https://doi.org/10.1515/icame-2015-0004.

Thomas, Lindsay. 'Modelling Long Novels: Network Analysis and A Brief History of Seven Killings'. *The Palgrave Handbook of Twentieth and Twenty-First Century Literature and Science*, edited by Neel Ahuja, Monique Allewaert, Lindsey Andrews, Gerry Canavan, Rebecca Evans, Nihad M. Farooq, and Erica Fretwell, Palgrave Macmillan, 2020, pp. 653–67, https://doi.org/10.1007/978-3-030-48244-2_37.

Thomson, Rachel, Janet Holland, Sheena McGrellis, Robert Bell, Sheila Henderson, and Sue Sharpe. 'Inventing Adulthoods: A Biographical Approach to Understanding Youth Citizenship'. *The Sociological Review*, vol. 52, no. 2, 2017, https://doi.org/10.1111/j.1467-954X.2004.00466.x.

Toolan, Michael. 'How Children's Literature Is Translated: Suggestions for Stylistic Research Using Parallel Corpora'. *Ilha Do Desterro*, vol. 71, 2018, pp. 151–67.

Trites, Roberta Seelinger. *Disturbing the Universe: Power and Repression in Adolescent Literature*. University of Iowa Press, 1998.

---. 'The Harry Potter Novels as a Test Case for Adolescent Literature'. *Style*, vol. 35, no. 3, 2001, pp. 472–85.

Underwood, Ted. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019.

Underwood, Ted, David Bamman, and Sabrina Lee. 'The Transformation of Gender in English-Language Fiction'. *Journal of Cultural Analytics*, vol. 3, no. 2, 2018.

Unicef. *Convention on the Rights of the Child*, 1989.

Uzuner, Özlem, and Boris Katz. *A Comparative Study of Language Models for Book and Author Recognition*. Springer, 2005, pp. 969–80, https://doi.org/10.1007/11562214_84.

van Dalen-Oskam, Karina, and Joris van Zundert. 'Modelling Features of Characters: Some Digital Ways to Look at Names in Literary Texts'. *Literary and Linguistic Computing*, vol. 19, no. 3, 2004, pp. 289–301, https://doi.org/10.1093/llc/19.3.289.

Van De Velde, Lieve. 'Hoe Kijkt de Wereld Naar Oudere Vrouwen?' *De Standaard*, 25 June 2022.

van den Hoven, Peter. 'Grensverkeer'. *Raster*, vol. 56, 1991, pp. 142–61.

---. 'Kijk, Volwassenen, Hoe Aardig Ook, Zijn Natuurlijk Wèl Tuig van de Richel. Peter van Den Hoven in Gesprek Met Guus Kuijer.' *BZZLLETIN*, vol. 73, 1980, pp. 46–60.

van Hulle, Dirk, and Mike Kestemont. 'Periodizing Samuel Beckett's Works: A Stylochronometric Approach'. *Style*, vol. 50, no. 2, 2016, pp. 172–202.

van Leeuwen, Joke. *Ik Heet Reinier En Ons Huis Is Afgebrand* (boekvoorstelling). online platform, 2020.

van Lierop-Debrauwer, Helma. 'Crossing the Border: Authors Do It, but Do Critics? The Reception of Dual-Readership Authors in the Netherlands'. *Transcending Boundaries: Writing for a Dual Audience of Children and Adults*, edited by Sandra L. Beckett, Garland, 1999, pp. 3–12.

---. 'Over de "Grote Gelijkenis": Adolescentenromans Voor Jongeren En Voor Volwassenen'. *Literatuur Zonder Leeftijd*, vol. 14, 2000, pp. 336–51.

van Lierop-Debrauwer, Helma, and Neel Bastiaansen-Harks. *Over Grenzen: De Adolescentenroman in Het Literatuuronderwijs*. Eburon, 2005.

Vanhees, Claudio, Mathea Simons, and Vanessa Joosen. 'Desirability of Multimedia Hyperlinks in Fiction to Foster Pupil Reading Motivation and Immersion: Reading Tools for MALL'. *Advances in Educational Technologies and Instructional Design*, edited by Alberto Andujar, IGI Global, 2020, pp. 79–109, https://doi.org/10.4018/978-1-7998-1097-1.ch004.

Villar, Feliciano, and Josep Fabà. 'Draw a Young and an Older Person: Schoolchildren's Images of Older People'. *Educational Gerontology*, vol. 38, no. 12, 2012, pp. 827–40, https://doi.org/10.1080/03601277.2011.645445.

Wall, Barbara. *The Narrator's Voice: The Dilemma of Children's Fiction*. Macmillan, 1991.

Walsh, John A. 'Comic Book Markup Language: An Introduction and Rationale'. *Digital Humanities Quarterly*, vol. 6, no. 1.

Wanner, Franz, Johannes Fuchs, Daniella Oelke, and Daniel A. Keim. 'Are My Children Old Enough to Read These Books? Age Suitability Analysis'. *Polibits*, no. 43, 2011, pp. 93–100.

Warner, Marina. *From the Beast to the Blonde: On Fairy Tales and Their Tellers*. Vintage, 1994.

Watson, Victor. 'Series Fiction'. *International Companion Encyclopedia of Children's Literature*, edited by Peter Hunt, 2nd ed., Routledge, 2004, pp. 532–41.

Westman, Jack C. 'Juvenile Ageism'. *Dealing with Child Abuse and Neglect as Public Health Problems*, Springer, 2019, pp. 39–48.

Whited, Lana, editor. *The Ivory Tower and Harry Potter: Perspectives in a Literary Phenomenon*. University of Missouri Press, 2002.

Wilson, Jacqueline. *Interview with Jacqueline Wilson*. Interview by Vanessa Joosen, 22 Oct. 2020.

Woltz, Anna. *Interview with Anna Woltz*. Interview by Vanessa Joosen, 13 Jan. 2021.

Woodward, Kathleen. 'Against Wisdom: The Social Politics of Anger and Aging.' *Cultural Critique*, vol. 51, 2002, pp. 186–218.

Woodward, Kathleen, editor. *Figuring Age: Women, Bodies, Generations*. Indiana University Press, 1999.

Wybo, S. 'Ik Kruip in de Huid van Mijn Personages'. *Het Laatste Nieuws*, 19 Feb. 2002, p. 14.

Xu, Liyan, and Jinho D. Choi. *Online Coreference Resolution for Dialogue Processing: Improving Mention-Linking on Real-Time Conversations*, 2022, https://doi.org/10.48550/ARXIV.2205.10670.

Yi Ma, Eric Fosler-Lussier, and Robert Lofthus. 'Ranking-Based Readability Assessment for Early Primary Children's Literature'. *Human Language Technologies*, Association for Computational Linguistics, 2012, pp. 548–52.

Yu, Bei. 'An Evaluation of Text Classification Methods for Literary Study'. *Literary and Linguistic Computing*, vol. 23, no. 3, 2008, pp. 327–43, https://doi.org/10.1093/llc/fqn015.

Yule, G. Udny. 'On Sentence-Length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship'. *Biometrika*, vol. 30, no. 3/4, 1939, pp. 363–90.

Zhu, Yukun, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 'Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books'. *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 19–27.

Appendices

Appendix A: annotation manual

A1 Setup

A1.1 List of documents

This is a list of the documents you receive at the start of your annotation assignment:

- **annotation file** (.xml): contains the header and the text of the book you will be annotating, the name of this file is formatted as follows: AUTHOR_title_year.xml (e.g., PULLMAN_northernlights_1995.xml)
- **character list** (.xlsx): empty apart from a table heading
- **schema** (.rng): contains all rules your annotated file should adhere to
- **logbook** (.docx): a questionnaire to fill out during or after the annotation process
- optional: **corpus file** (.xlsx): an overview of the CAFYR corpus that includes information needed to fill in the header
- optional: **original text** (.txt or .pdf): a plain text or pdf version of the book you will be annotating
- optional: **tester script** (.py): a python script interns and team members use to check their finished annotation for errors

A1.2 Oxygen XML Editor

Install version 20.1 of the program Oxygen XML Editor on your computer or laptop. If you are a student or intern working on the CAFYR project, you will receive a license key for this program. Do not share this key with anyone. You can also annotate your file in Sublime Text but you need Oxygen XML Editor to regularly check your file for errors.

To link your annotation file to the CAFYR schema, follow these steps:

4. open your annotation file in Oxygen XML Editor
5. navigate to Document > Schema > Associate Schema
6. click the folder icon in the top right corner of the pop-up screen
7. navigate to your local folder containing 'CAFYRschema.rng'
8. click 'Open', then 'OK'
9. check your document for errors by clicking the 'Validate' icon in the toolbar (the icon looks like a page with a red tick across it)

If any errors or problems show up at this stage, contact PhD1.

Note that it is best not to move the schema file or your annotation file while working on it as the referencing works with an absolute path. If you do move either of these files, you will need to re-associate your schema.

A2 Header

The TEI header contains information about the annotation process and the original book. The sections that need to be filled out are marked with 'xxx'. Do not change or delete any of the other sections in the header unless this guide states otherwise. Examples on what needs to be entered for every line are in the document itself and a detailed overview can be found below.

A2.1 Title and publication

In this section, we record information about the annotation process.

```
<title xml:id="xxx">xxx</title>

<title xml:id="PULLMAN_northernlights_1995">PULLMAN_northernlights_1995</title>
```

Enter the title twice, both after 'id' and between the title-tags. This should match the name of your document (without .xml) and consist of the author's last name (in all caps), the title of the book (all lowercase) and the year of first publication, separated by underscores.

```
<author ref="#xml:idxxx">xxx</author>

<author ref="#xml:idIM">Ian McEwan</author>
```

Enter the author's initials (in capitals) after #xml:id and enter the author's full name between the author-tags.

```
<name>xxx</name>

<name>Vanessa Joosen</name>
```

Enter your own name.

```
<date>xxx</date>

<date>19/10/2018</date>
```

Enter the final date of annotation. You'll have to fill this out after you've finished the annotation.

A2.2 Series and source

In this section of the header, you enter information that relates to the orignal text.

```
<seriesStmt>
<title>xxx</title>
<biblScope unit="volume">xxx</biblScope>
</seriesStmt>

<seriesStmt>
```

```
<title>Harry Potter</title>
<biblScope unit="volume">1</biblScope>
</seriesStmt>
```

First check if your book is part of a series. If it is, enter the series info. If it is not, delete this section instead of filling it out.

```
<title>xxx</title>

<title>Harry Potter and the Philosopher's Stone</title>
```

Enter the title of the book and, if applicable, the subtitle.

```
<author key="#xml:idxxx">xxx</author>

<author key="#xml:idIM">Ian McEwan</author>
```

Enter the author's information again. This should correspond to the info about the author in the title statement.

```
<editor>xxx</editor>

<editor>John Smith</editor>
```

If an editor if listed, enter his/her name here. If not, delete this line.

```
<respStmt>
<resp>illustrator</resp>
<name>xxx</name>
</respStmt>

<respStmt>
<resp>illustrator</resp>
<name>Joke van Leeuwen</name>
</respStmt>
```

If the book contains illustrations, fill in the illustrator's name. If you cannot find it, write 'unknown' instead of a name. If there are no illustrations, delete this section.

```
<date>xxx</date>

<date>2017</date>
```

Use the Excel file 'Corpus ERC' to enter the year of first publication. If this information is not included in the file, use the Dutch Centraal Bestand Kinderboeken (Central Database of Children's Books). This holds entries for most English books and all books in Dutch.

```
<pubPlace>xxx</pubPlace>

<pubPlace>Amsterdam</pubPlace>
```

Check the Centraal Bestand Kinderboeken to enter the place of first publication. This information can often be found by looking up the city of the headquarters of the original publisher.

```
<idno type="ISBN">xxx</idno>

<idno type="ISBN">9780751563603</idno>
```

Enter the ISBN number (preferably the ISBN-13) without hyphens or whitespace. You can often find this in the original text document.

A2.3 Author

The following sections of the header contain more details about the author.

```
<person xml:id="xxxx">
<persName>xxx</persName>

<person xml:id="JKR">
<persName>J.K. Rowling</persName>
```

Make sure this matches the information that you have already filled in above (twice).

```
<birth>xxx</birth>

<birth>1952</birth>
```

Look up the author's year of birth in the Excel file 'Corpus ERC'. If this information is not included in the file, look it up on the internet (use only reliable sources).

```
<age when="0001" value="0">xxx at the date of publication (xxx)</age>

<age when="2018" value="41">41 at the date of publication (2018)</age>
```

Look up this information in the Excel file 'Corpus ERC'. If this information is not included in the file, calculate the age of the author in the year of publication. We will not be looking at the specific month in which a book was published, or an author had their birthday, so simply calculate this as follows: year of publication minus year of birth (e.g., 2018 - 1977 = 41). Note that you have to fill in this information in the tag text as well as in the attribute values (replace '0001' and '0').

- 'when' refers to the date of publication
- 'value' refers to the age of the author at that time

```
<sex value="xxx">xxx</sex>

<sex value="F">female</sex> or <sex value="M">male</sex>
```

Enter the author's sex (in a binary system)

```
<nationality key="xxx">xxx</nationality>

<nationality key="UK">British</nationality>
```

Look up this information in the Excel file 'Corpus ERC'. For the author's nationality, we use country codes BE (Belgisch), NL (Nederlands), and UK (British).

A2.4 Various

Finally, we record some more metainformation about the book.

```
<age>xxx</age>
```

```
<age>7</age>
```

Enter the age of the book's intended reader. This reader as labelled as 'conceptual' because it does not refer to a real person. To determine the intended reader's minimal age you first check the Excel file 'Corpus ERC'. If this information is not included in the file, check your book. If there is an age (e.g., 8+) mentioned, use that one (8). If not, look up your book in Centraal Bestand Kinderboeken. If there is an age range listed, use the lowest number that you find there.

```
<language ident="xxx"></language>
<language ident="ENG"></language>
```

Enter either "ENG" for English books or "NL" for books in Dutch.

```
<term>xxx</term>
<term>fantasy</term>
```

There are two term-tags to fill in. In the first one, you enter the book's genre by choosing one of the following: realism, fantasy, historical novel, adventure. Determine this by checking the Centraal Bestand Kinderboeken or judge the genre yourself after finishing the annotation.

```
<term>xxx</term>
<term>early work</term>
```

In the second tag-set, you fill in the place of the book in the author's oeuvre. Check the Excel file 'Corpus ERC' to determine this. Here you choose one of the following: debut (for the first book), early work (first ten years after debut), middle work (11-30 years after debut), late work (30+ years after debut), final book (only for authors who have died).

A3 Text

On this page, you will find a step-by-step guide to annotating a text.

In this stage, we

1. identify and add information to passages of direct, indirect or reported speech
2. identify and add information to passages in texts that give a metareflection on a specific age or stage of life
3. identify all character references.

If you are not familiar with TEI markup language, please read the following segment on TEI basics. If you are familiar with annotating texts using TEI, you can continue on to the section on annotating speech.

A3.1 TEI basics

Why do we tag?

There are many digital tools that can analyse plain text, such as found in a .txt file. The information gathered from this basic form of digital texts, however, is relatively limited; for example, when a text refers to “London” without providing context, the computer analysing this text will not be able to distinguish between the capital of England and the city in Ontario with the same name. When we want to answer more specific questions about a text concerning details about characters, we first need to add an extra layer of information to that text.

How do we tag?

The added layer of information transforms a plain text to a **TEI document**.

> [TEI’]s purpose is to provide guidelines for the creation and management in digital form of every type of data created and used by researchers in the Humanities, such as source texts, manuscripts, archival documents, ancient inscriptions, and many others. (Burnard 2014)

To create a TEI document, the encoding language XML is used. This language formats features of a text, or word, as tags which indicate functions or semantics. For example, where a simple word processor (e.g., Microsoft Word) would display an underlined word like this, XML formats the same function like `<u>this</u>`.

- The symbols < and > mark the start and end of tags
- A tag can take one of two forms: a start-tag (<u>), which indicates the point at which an element starts, or an end-tag (</u>), which indicated the end of that element[i]

Tags are not used solely to encode layout, but also to add extra information to words, sentences, paragraphs, etc. One of the tags used in the CAFYR-project is the said-tag, which we use to identify fragments of text that are part of characters’ dialogue. When these tags are added to an entire text, digital tools can distinguish between the voice of the narrator and the speech instances of the fictional characters. Extra data may be added to a tag in the form of attribute specifications. Not only can a fragment of text be identified as the direct speech of a character, but an attribute can specify for example which character is speaking, which would look like this: `<said who="marie">I like coconuts</said>`.

- A tag always begins with a name, followed by optional attributes
- An attribute always has a value, in this case it is the name of the character who is speaking
- The end-tag never contains attributes

For more information about XML, visit the TEI-website and their page on using XML.

A3.2 Annotating direct, indirect and reported speech

The base text you receive for your annotation should already include tags for direct and indirect speech. Note that this preannotation may contain mistakes; it is only as good as the quality of

the digital text available. When you notice such a mistake, please correct it. Texts with first or third-person narration are annotated differently in terms of speech.

Third-person

In case of a story told from a third-person perspective, we distinguish direct from indirect speech. All text from the narrator is considered to be indirect speech, while text uttered by characters (most of the time marked by quotation marks) is tagged as direct speech. For example (Rowling 1997):

```
<said direct="false">His aunt was back outside the door.</said><said
direct="true" who="xxx">‘Are you up yet?’</said><said direct="false">she
demanded.</said><said direct="true" who="xxx">‘Nearly,’</said><said
direct="false">said Harry.</said>
```

Breakdown of this annotated passage:

- <said direct="false">: indication of indirect speech
- <said direct="true" who="xxx">
  - <said direct="true": indication of direct speech
  - who="xxx">: identity of the speaker, expressed with an id (see the page on compiling a character list)

First-person

In this case no indirect speech is tagged. Consider all descriptions as direct speech from the narrator. Use the id ‘narrator’ for these tags even if the narrator is a character in the story. Add the id to the character list; all information of the narrating character may be copied to the id ‘narrator’. Direct speech of this character is to be tagged with the id of the character itself as are all references to this character. For example (Van Leeuwen 2010):

```
<said direct="true" who="narrator">At that time <rs ref="toda">I</rs> never
thought it was somewhere else. It was always somewhere else, except where <rs
ref="toda dad">we</rs> lived. Everyone could pronounce <rs ref="toda">my</rs>
name without difficulty.</said>[ii]
```

As you go through the text, change xxx to the correct character id (see compile a character list). In case there are multiple characters talking simultaneously, it is possible to tag more than one speaker, like this: <said direct="true" who="fred george">. In some cases, you will have to manually add tags for direct speech, for example when a letter is included in the story. Letters are considered to be direct speech of the character who wrote the letter.

Reported speech

Reported speech is more difficult to tag automatically and has to be manually added by the annotator. Any passage that paraphrases something said by a different character than the one speaking (or different from the narrator in the case of indirect speech) is considered to be reported speech. For this type of speech we use a <said> tag including the attribute @ana with the value #reported. For example (Rowling 1997):

```
<said direct="true" who="draco">‘Think my name’s funny, do you? No need to ask
who you are. My father told me <said ana="#reported" who="lucius">all the
Weasleys have red hair, freckles and more children than they can
afford.</said>’</said>
```

Note that the reported speech tag is enclosed in the direct speech tag of ‘draco’. The first end tag corresponds to the reported speech tag and the second to the direct speech tag.

A3.3 Annotating metareflections on age or life stage

Metareflections are clear statements given by a narrator or character, for example:

> “Children do not have a house of their own yet,” the doorman said. “It belongs to their parents.” (Van Leeuwen 2013)[iii]

The information that needs to be added to this type of passage is the following: which specific age or life stage is being discussed?

```
<said direct=”true” who=”doorman”><seg about=”child”>“Children do not have a
house of their own yet,”</seg></said>
```

Breakdown of this annotated passage:

- `<seg about=”child”>`: indication of age/life stage discussed
  - `child`: id of the age/life stage according to the age model
- `</seg></said>`: closing tags

A3.4 Annotating character references

In our annotation method all references to characters need to be tagged with their corresponding character. These references can take the shape of:

- Characters’ names (first, last, full, nicknames, …)
- Pronouns
- Nouns that refer to characters, such as boy / girl, aunt / uncle, people, … (tag only the noun without its unit or possible adjectives)

To identify these character references we us the `<rs>` tag (rs stands for referencing string in TEI), including the attribute `@ref`. As you work your way through the text, change the value of `@ref` to the correct character id. When we add this stage of the annotation process to the example given for direct and indirect speech, it looks as follows (Rowling 1997):

```
<said direct="false"><rs ref="harry10">His</rs> <rs ref="petunia">aunt</rs> was
back outside the door.</said>
<said direct="true" who="petunia">‘Are <rs ref="harry10">you</rs> up
yet?’</said>
<said direct="false"><rs ref="petunia">she</rs> demanded.</said>
<said direct="true" who="harry10">‘Nearly,’</said>
<said direct="false">said <rs ref="harry10">Harry</rs>.</said>
```

In case there is a reference to multiple characters (they, their, we, …), it is possible to tag more than one character, like this: `<rs ref="vernon petunia">Dursleys</rs>`. Do this only for a maximum of three characters, e.g., not an entire family.

In the pre-annotated version of your text, most pronouns are automatically tagged (with the default character value 'xxx'). If you come across other pronouns as well as characters' names and nouns referring to characters, you can use the Find and Replace function in Oxygen to tag all instances of these words/names. Make sure F+R is set to 'whole words only' before executing the following process:

Find: `yeh` Replace: `<rs ref="xxx">yeh</rs>`

Find: `Harry` Replace: `<rs ref="harry">Harry</rs>`

Note that not all pronouns or nouns refer to a character in the story; they can also refer to objects or very fleeting side-characters that are not worth tagging. You may have to delete several tags.

Special cases

Je: this special case occurs most often in Dutch texts, where 'je' is used as a general pronoun. In these instances, 'je' does not need to be tagged. (e.g., 'Dat deed je ook niet als je net ontsnapt was.' and 'Honden huilen verderop in het veld en katten janken naast je voordeur.') This also occurs in English with 'you', although it seems to be less frequent.

Professions or other characterizing groups: you do not need to tag the profession of a character nor the social group he or she belongs to (e.g., do not tag 'builder' in the following sentence: Bob was a builder and he had his own tv-show.)

A4 Character list

The character list is used to keep track of fictional characters several of their features, for example, species, gender, and ethnicity. To make sure you do not forget to add characters to the file, it is good practice to enter the data while annotating your book. The file consists of twelve table headings which each correspond to a feature.

A4.1 Setting up your file

Change the name of the Excel file 'CAFYRcharacterlist.xlsx' to match the name of your annotation file (format: AUTHOR_title_year), for example:

ROWLING_harrypotterandthephilosophersstone_1997.xlsx
VANLEEUWEN_iep_1996.xlsx

A4.2 Age model

Numerical ages will not be specified for all the characters. To make sure that we annotate vaguer descriptions consistently, we all follow the same age model for annotating characters' ages.

Make sure to be alert for shifts in a character's age, because of time lapses or the progression of time. Life stages can be divided in different ways and linked to various numerical ages. To be consistent here, we use the following model, based on Lorraine Green (2010) and Thomas Armstrong (2007).

Life stage	Corresponding numerical age
unborn	
infant	0–2
child	3–11
● earlychild	3–5
● middlechild	6–8
● latechild	9–11
adolescent	12–19
adult	20–…
● earlyadult	20–39
○ twenties	20–29
○ thirties	30–39
● middleadult	40–59
○ forties	40–49
○ fifties	50–59
● oldadult	60–79
○ sixties	60–69
○ seventies	70–79
● deepoldadult	80–…
○ eighties	80–89
○ nineties	90–99

Note: it is possible that a character is referred to in a very general way (e.g., "He has always been mum's favorite." "When my father was 10 years old.") In this case you can use an ageless version of the character to which you assign the life stage 'ageless').

Note: animal characters should only be tagged when they are anthropomorphized in the story. To determine their age you can only choose from the following life stages: 'unborn', 'infant', 'child', 'adult' and 'oldadult'.

A4.3 Entering data

Each column (apart from the first, 'id') should only be filled in when the information is clearly stated in the text. Each feature can be filled out as 'not mentioned' in case the information is not given or if you are not sure about something. Note that some of the information will not be available to you the first time a character appears in the text and that you may have to come back to the list to update it.

If you come across a character that belongs in a category that is not listed in the drop-down menu but you think it is relevant to be more specific in light of the research, please consult the PI or PhD1.

Id

When a character appears for the first time, as either a speaker or a subject, create an id in the first column. The id consists of only one string, all lower case, without whitespaces: in the case of an individual character this is preferably the name by which the character is best known. This will also be the id with which you tag this character in the text. It can be:

- The character's first name (e.g., harry, dudley)
- The character's last name (e.g., dumbledore). Be careful: if you choose a character's last name, make sure it can be distinguished from other characters (e.g., do not use "weasley" for Ron's father, but "mrweasley" or "arthur").
- A group name (e.g., students). Use this in case the character you want to add is a group. Keep in mind that not all groups are worth tagging; when the composition of a group is very diverse (including more than one life stage, e.g., a family) you can disregard it. When the group has a specific age range (e.g., students) it is relevant to our research.
- One of the options above + age (e.g., harry10). In case the age of the character changes in the narrative you should use this id-format together with creating various entries for that character in the character list. For example, in Harry Potter and the Philosopher's Stone, Harry appears as a one-year-old child, a ten-year-old boy, an 11-year-old boy, and an adolescent. We create four id's for Harry: harry1, harry10, harry11, harryadolescent. See the age model for available options.
- After compiling the entire character list of your annotated book, sort the columns alphabetically according to id. Do this by selecting a random cell containing an id and then using the 'Sort A-Z' function in Excel.

Full name

The character's full name, first and last name capitalized, belongs in the second column. If the full name of the character is not stated, leave this cell open.

Individual / group

This column is formatted as a drop-down menu from which you can choose 'individual' or 'group'. In the case of a group, specify the common features in the following cells and put the rest down as 'other'.

Species

This column is formatted as a drop-down menu from which you can choose 'human', 'animal', 'animated object' or 'other'. Animals should only be tagged when they are anthropomorphized in the story. We use a simplified version of the age model to determine the age of such a character (check the age model).

Gender

This column is formatted as a drop-down menu from which you can choose 'male', 'female', 'other' or 'not mentioned'.

Ethnicity

This column is formatted as a drop-down menu from which you can choose ‘white’, ‘black’, ‘asian’, ‘native american’, ‘north african’, ‘biracial’, ‘other’, ‘n/a’ or ‘not mentioned’. The category ‘n/a’ (not applicable) is used when the character is an animal, animated object or other and no ethnicity is specified.

Relation to protagonist

These columns are formatted as drop-down menus from which you can choose ‘protagonist’, ‘adversary’, ‘family’, ‘parent’, ‘sibling’, ‘child’, ‘friend’, ‘helper’, ‘teacher’, ‘ambivalent’, ‘other’ or ‘none’. As there can be only one protagonist, you might have to select one character as main protagonist when your story features multiple protagonists. The column ‘relation to protagonist 2’ is optional.

Age and life stage

Age and life stage are special cases; a character may grow older in the course of the story, and/or the narrative may contain flashbacks or flash-forwards that involve a shift in age. Since age is central to the project, we want to trace it as precisely as possible.

For each entry of each character (in case of a change of age in the book) you should either specify its numerical age (0–100) or its life stage. The column life stage is formatted as a drop-down menu from which you can choose the correct life stage (see the age model below for available options).

Certainty

This column is formatted as a drop-down menu from which you can choose the certainty with which you assigned the numerical age or life stage to the character. There are two possible levels:

- High: when age or life stage is explicitly mentioned
- Medium: when you state the average of a limited range (e.g., when it is mentioned that a character very recently graduated from secondary school, he or she is between 17 and 19 years of age, so you tag this as age=”18” with a medium certainty) or when you take an educated guess towards a life stage (e.g., characters with a profession are assumed to be adults)

Notes

The last column, titled ‘notes’, is a free cell in which to write down your thought process when assigning a numerical age or life stage. For example, when a character’s birthday is mentioned, create a new entry with a unique id and add a note along the lines of “11th birthday mentioned”.

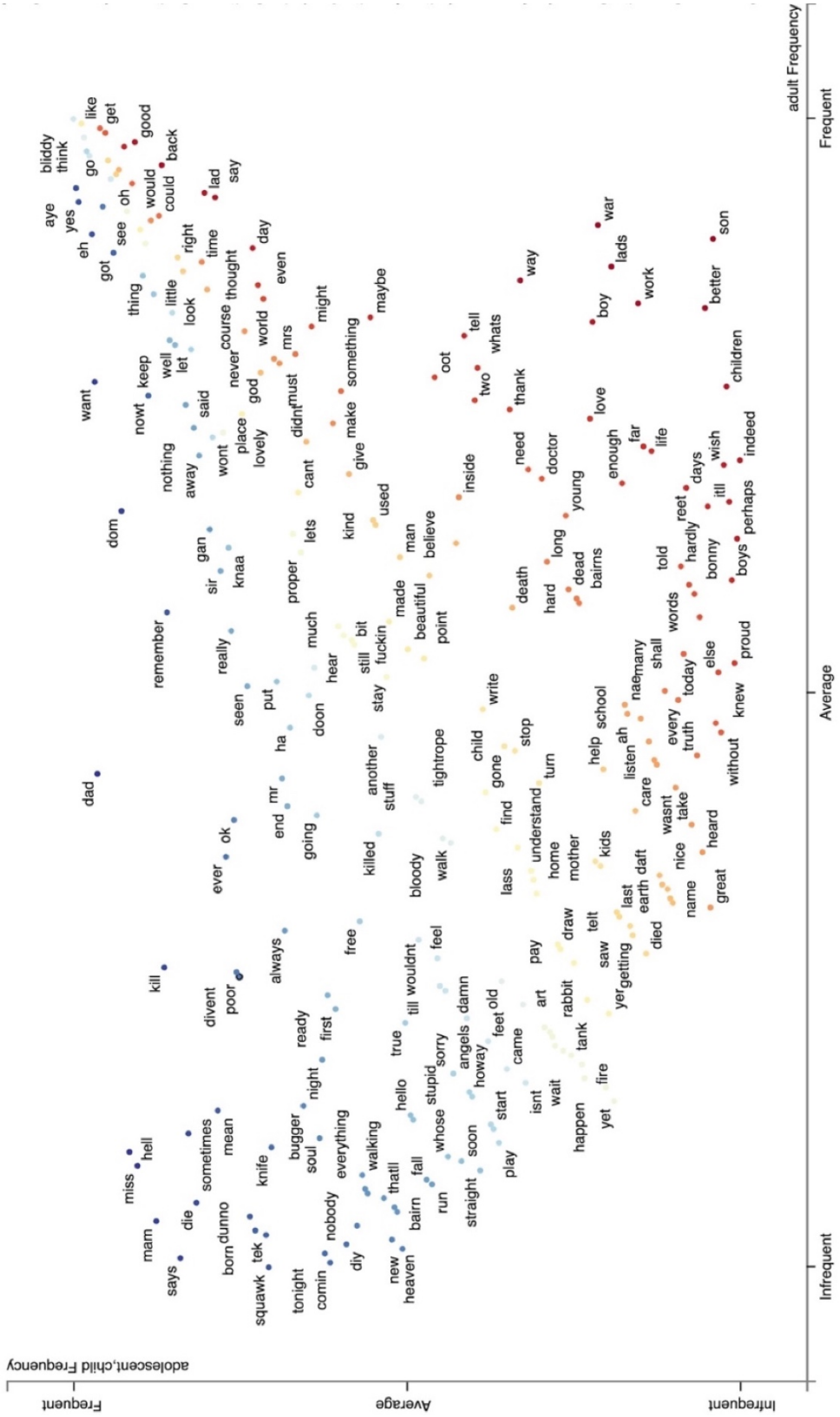
[i] A third kind of tag, the composite tag, merges these two forms.

[ii] My translation, original text: “Ik dacht in die tijd nooit dat het daar ergens anders was. Het was overal ergens anders, behalve waar we woonden. Iedereen kon zonder moeite mijn naam uitspreken.”

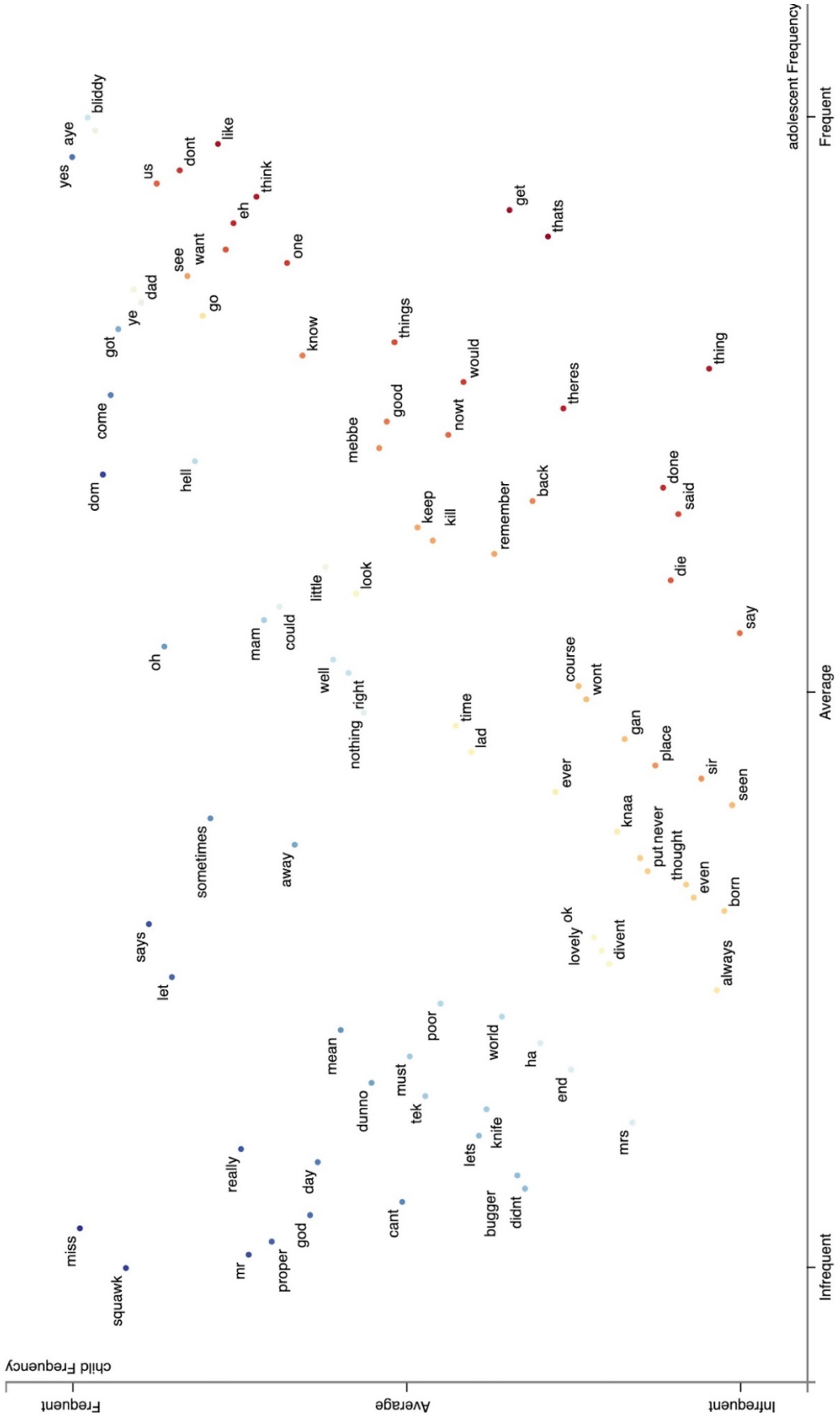
[iii] My translation, original text: “Kinderen hebben nog geen eigen huis,” zei de portier. “Dat is van hun ouders.”

Appendix B: scatterplots of character speech in the oeuvre of David Almond, categorised according to the age of the intended reader

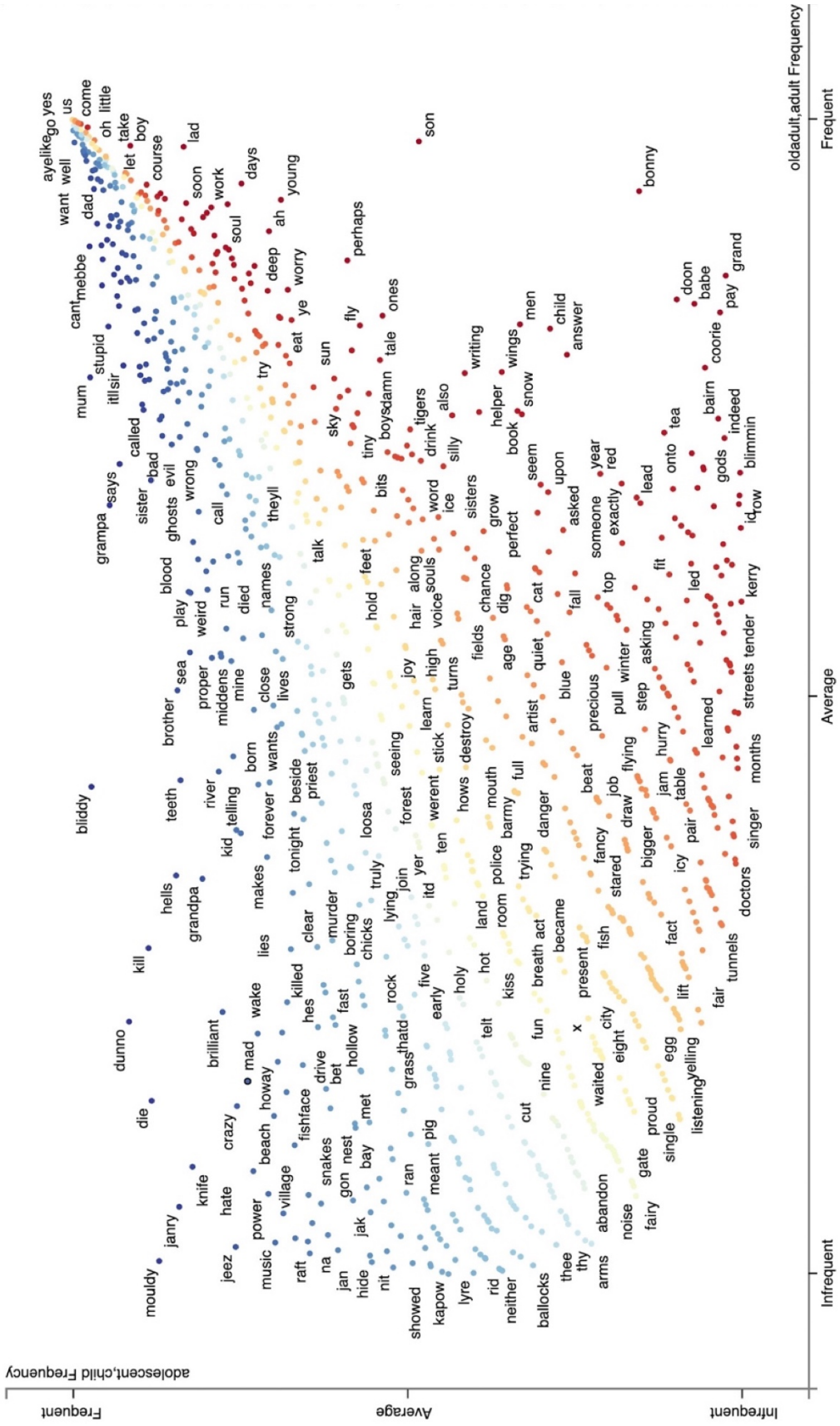
1. Scatterplot comparing speech of child and adolescent characters to speech of adult characters in David Almond's novel for adults, after removal of stop words and character names.



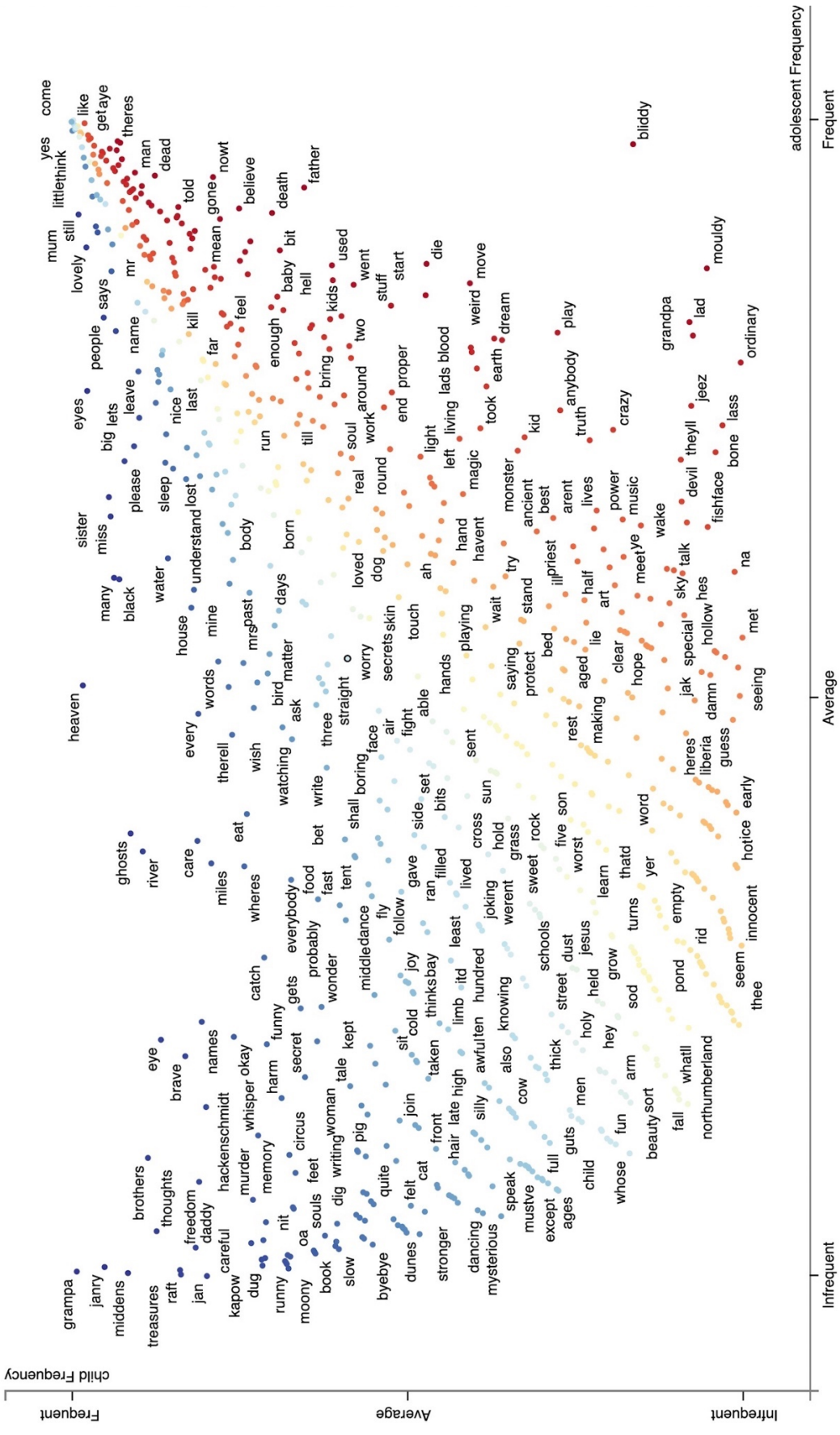
2. Scatterplot comparing speech of child characters to speech of adolescent characters in David Almond's novel for adults, after removal of stop words and character names.



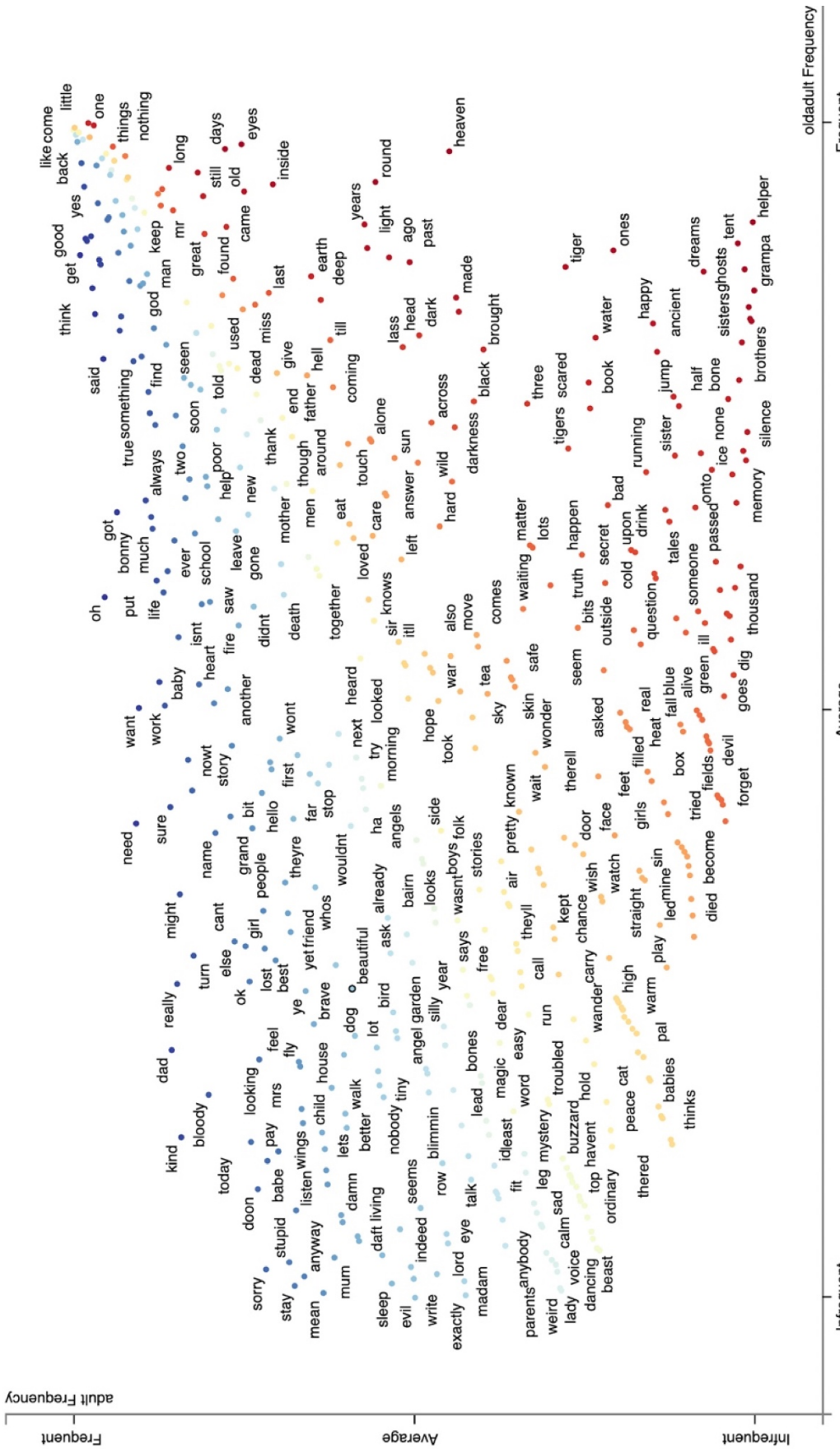
3. Scatterplot comparing speech of child and adolescent characters to speech of adult and oldadult characters in David Almond's novels for young adults (aged twelve to fifteen), after removal of stop words and character names.



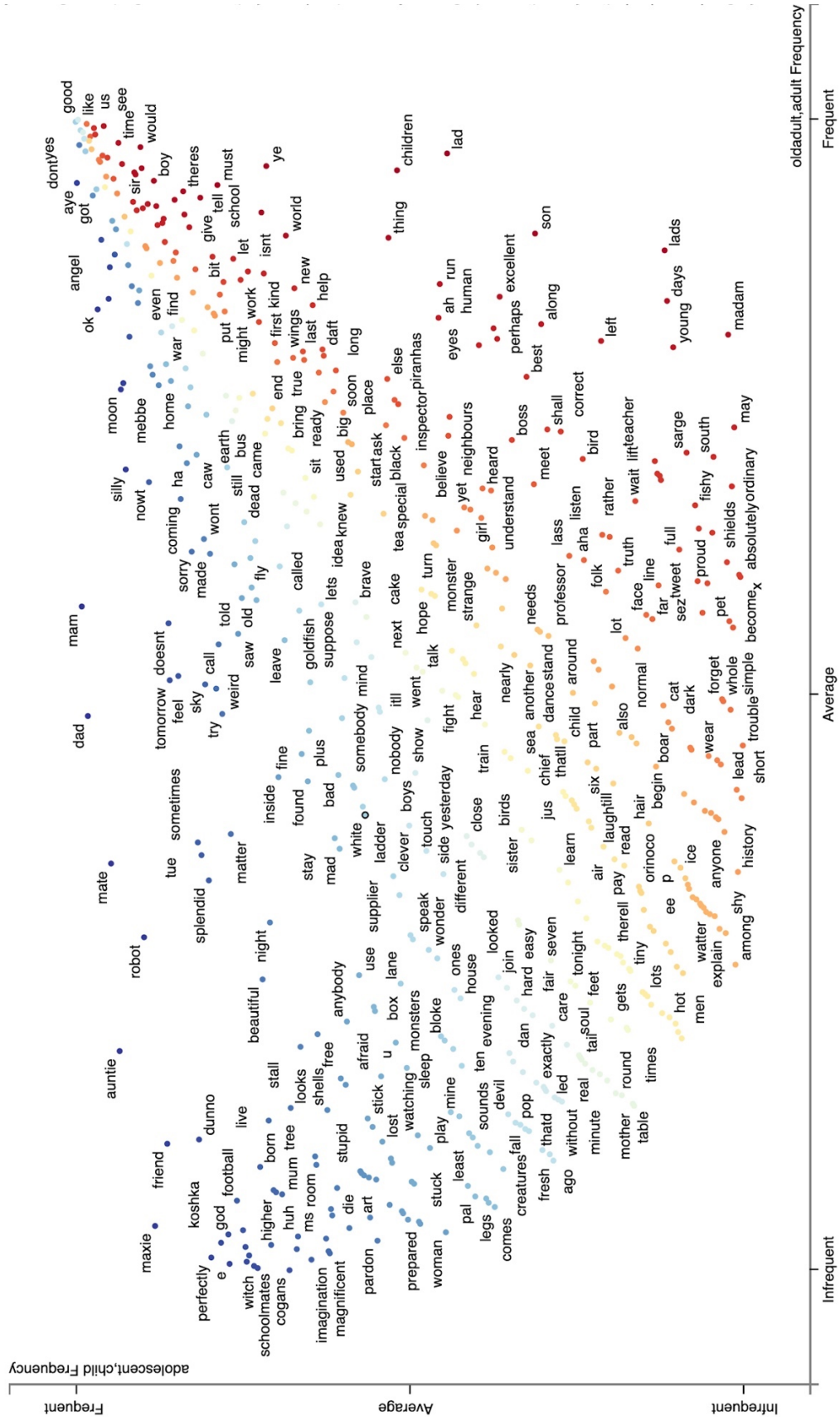
4. Scatterplot comparing speech of child characters to speech of adolescent characters in David Almond's novels for young adults (aged twelve to fifteen), after removal of stop words and character names.



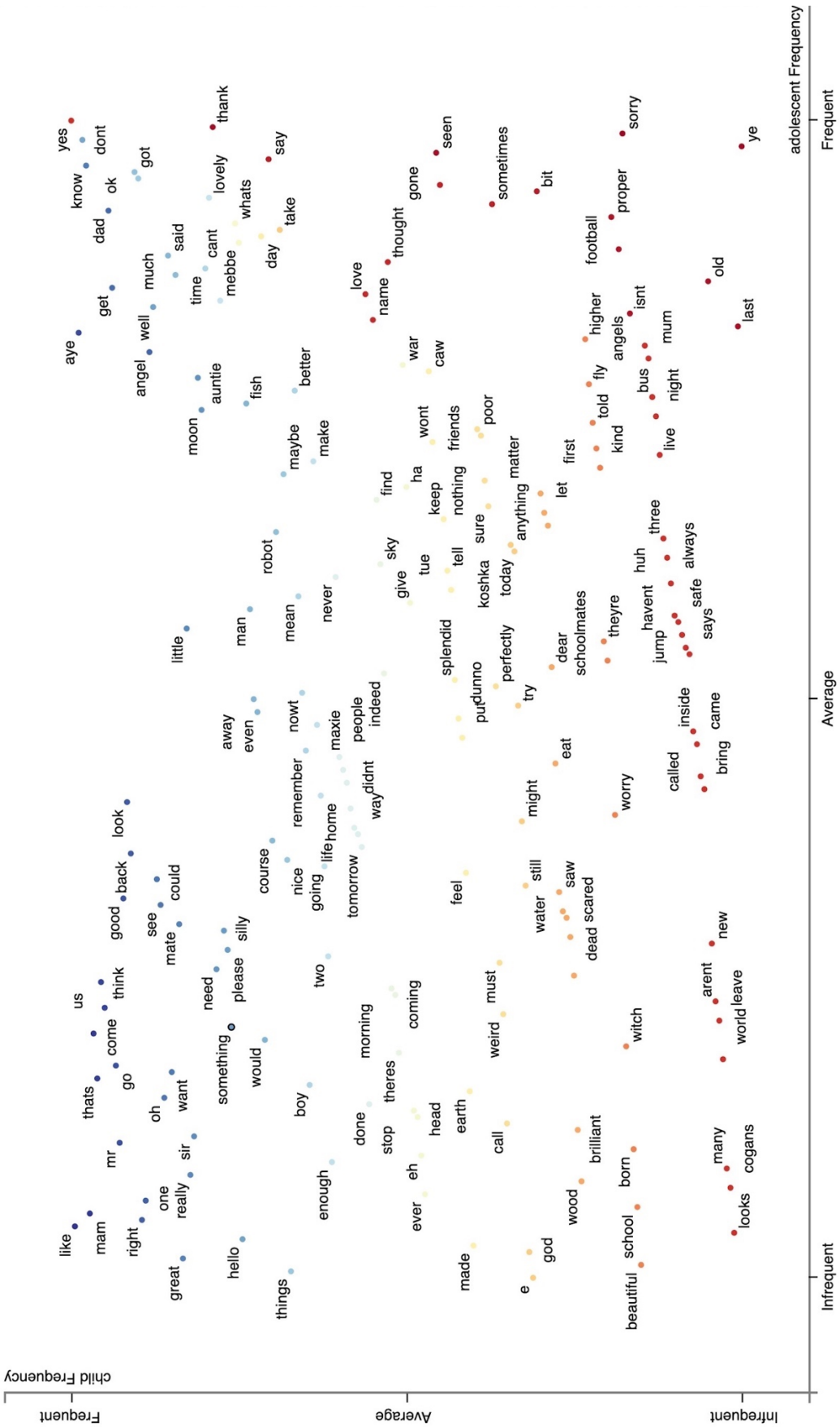
5. Scatterplot comparing speech of adult characters to speech of oldadult characters in David Almond's novels for young adults (aged twelve to fifteen), after removal of stop words and character names.



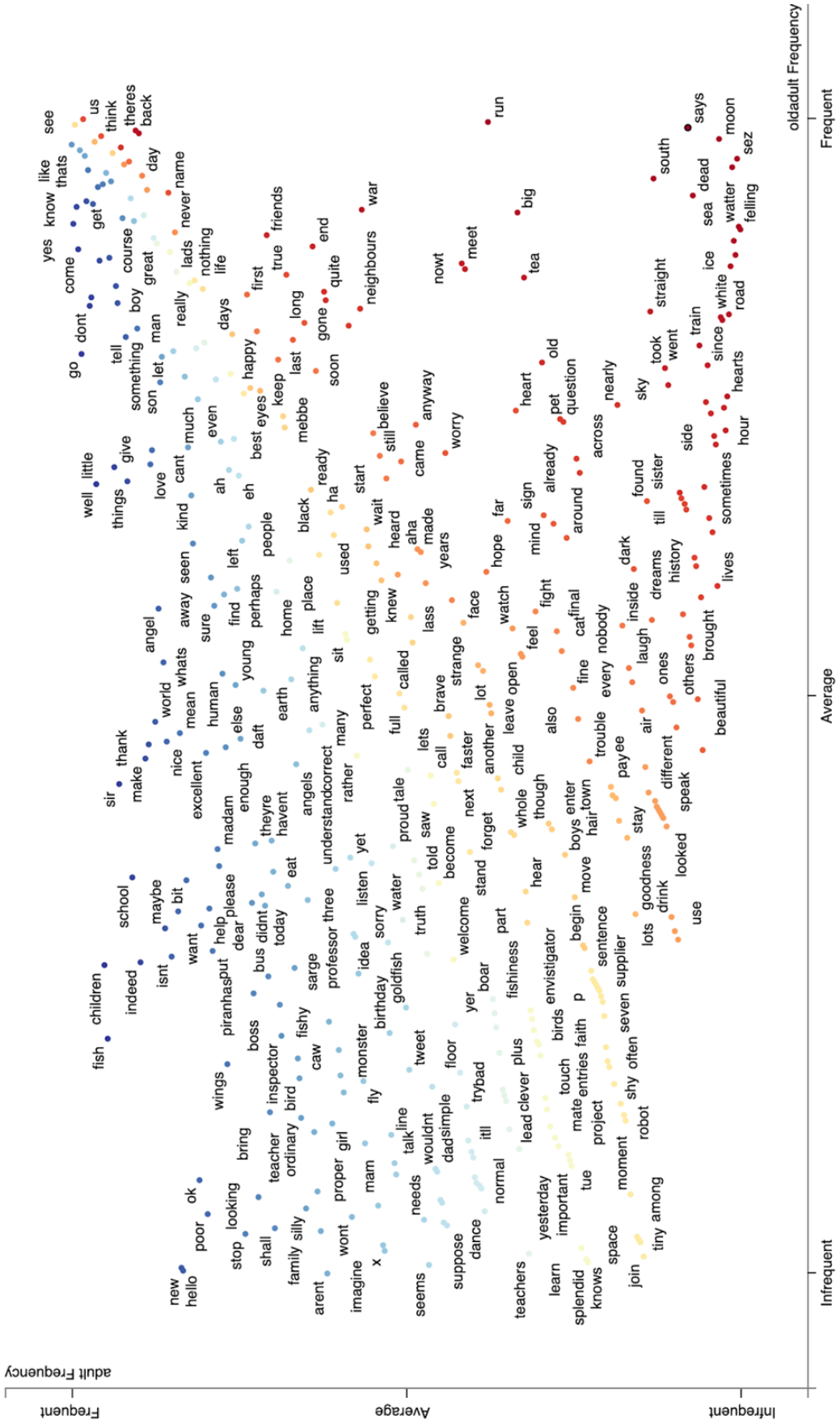
6. Scatterplot comparing speech of child and adolescent characters to speech of adult and oldadult characters in David Almond's novels for children (aged seven to nine), after removal of stop words and character names.



7. Scatterplot comparing speech of child characters to speech of adolescent characters in David Almond's novels for children (aged seven to nine), after removal of stop words and character names.



8. Scatterplot comparing speech of adult characters to speech of oldadult characters in David Almond's novels for children (aged seven to nine), after removal of stop words and character names.



Appendix C: list of 40 most common verbs, possessions and adjectives associated with child, adolescent, adult and oldadult characters for David Almond's oeuvre, grouped by the age of the intended reader

1. List of 40 most common verbs

intended readership	characters' age	terms
adult (18+)	child	say (77), see (24), think (22), do (22), come (19), go (19), have (18), draw (17), turn (14), want (13), get (13), look (12), know (12), yell (10), ' (10), take (10), hear (10), tell (10), whisper (9), make (9), walk (9), love (8), stand (8), 're (8), like (7), try (7), 'm (6), watch (6), laugh (6), hold (6), dream (6), begin (5), cry (5), mean (5), sit (5), read (5), use (5), giggle (5), keep (5), give (5)
	adolescent	say (376), do (71), think (60), have (55), get (54), see (54), know (54), laugh (43), go (43), take (38), whisper (35), want (33), look (29), hear (27), put (26), tell (26), hold (23), turn (22), come (21), try (20), remember (20), feel (19), 're (19), 'm (19), shrug (18), ' (17), answer (17), stand (16), watch (16), yell (15), breathe (15), run (14), keep (13), start (13), read (13), dream (12), ask (11), kill (11), write (11), point (10)
	adult	say (507), come (59), laugh (44), do (42), take (32), tell (32), have (31), ' (28), know (28), look (26), go (25), put (25), get (25), think (23), ask (23), 'm (23), whisper (23), turn (22), see (22), stand (19), smile (19), sigh (19), hold (18), yell (18), raise (14), keep (14), walk (14), move (13), give (13), speak (13), lift (12), lean (12), grin (12), call (12), cough (12), touch (11), murmur (10), light (10), shrug (10), want (10)
	oldadult	die (1), sit (1)
young adult (12–15)	child	say (1386), see (246), know (234), look (202), whisper (201), think (194), go (175), have (148), tell (145), come (143), do (138), want (113), feel (111), hold (96), take (94), hear (91), turn (84), laugh (79), keep (74), get (72), sit (71), put (70), stand (66), smile (63), find (60), ' (60), try (58), write (57), watch (57), stare (56), 'm (53), giggle (52), ask (51), imagine (51), walk (49), touch (49), remember (48), start (48), call (46), 're (45)
	adolescent	say (3394), see (599), know (573), look (481), think (453), go (441), whisper (420), do (417), want (321), tell (318), come (291), feel (282), have (261), hear (260), laugh (234), get (225), take (209), 'm (206), 're (198), turn (188),

		hold (174), ’ (167), stand (165), try (154), put (150), watch (133), keep (130), walk (123), find (121), shrug (119), stare (119), shake (114), ask (114), make (114), imagine (110), move (108), start (106), smile (105), sit (104), play (95)
	adult	say (2067), come (227), look (172), tell (140), have (140), smile (139), hold (133), whisper (131), see (128), know (124), go (118), do (117), take (117), laugh (113), think (104), put (99), stand (92), ask (88), get (87), call (79), ’ (78), turn (76), ’m (72), want (71), reach (67), use (57), sigh (57), shake (56), yell (54), watch (54), make (49), keep (46), sit (46), give (45), lean (44), touch (42), open (42), stare (41), lift (41), raise (40)
	oldadult	say (241), whisper (51), smile (38), tell (30), hold (28), turn (24), stare (22), look (22), come (22), take (21), have (20), put (17), laugh (16), write (16), sit (15), see (15), go (15), ask (15), do (14), know (14), stand (14), lift (13), sigh (13), touch (11), raise (11), use (11), hear (11), dig (11), call (10), remember (10), murmur (10), lean (10), ’ (9), reach (8), close (8), get (8), die (8), find (8), press (7), think (7)
children (7–9)	child	say (1140), look (208), know (182), do (174), see (162), ’ (154), get (152), think (143), have (140), go (122), whisper (95), want (95), tell (81), come (78), take (75), ’m (74), stand (73), ’re (64), turn (62), hear (61), stare (59), put (56), ask (54), feel (53), laugh (50), sit (48), try (47), like (42), need (40), run (38), start (37), call (37), walk (36), smile (34), move (33), keep (33), yell (33), hold (32), shake (31), reach (30)
	adolescent	say (15), paint (9), ’m (5), see (5), do (4), think (4), become (3), have (3), know (3), disappear (2), outrun (2), grow (2), mutter (2), whisper (2), admit (2), gasp (2), ’re (2), make (1), tell (1), ’s (1), ’ (1), throw (1), attend (1), hide (1), re (1), nick (1), feel (1), want (1), pretend (1), shrug (1), hesitate (1), trot (1), watch (1), linger (1), seem (1), play (1), hear (1), gain (1), stare (1), lay (1)
	adult	say (42), laugh (4), smile (4), lift (4), tell (3), come (2), take (2), unfold (2), hold (2), make (2), ’ (2), sit (2), open (2), photograph (2), think (2), visit (1), find (1), let (1), bend (1), push (1), help (1), shrug (1), hear (1), link (1), rest (1), read (1), send (1), hammer (1), sez (1), go (1), toot (1), call (1), tug (1), grin (1), taste (1), watch (1), ponder (1), snap (1), trot (1), bring (1)

	oldadult	say (140), see (22), look (15), turn (13), have (11), take (11), know (11), reach (9), think (9), do (9), tell (9), sigh (8), go (8), stand (6), open (6), come (6), ’ (5), keep (5), point (5), smile (5), put (5), drive (5), ask (5), need (4), ’m (4), drink (4), shake (4), hold (4), whisper (4), tick (4), yell (3), laugh (3), sit (3), stare (3), close (3), drop (3), lean (3), grin (3), murmur (3), mean (3)

2. List of 40 most common possessions

intended readership	characters’ age	terms
adult (18+)	child	hand (30), eye (14), arm (11), dad (8), side (7), mother (6), name (6), finger (6), head (6), mam (5), shoulder (5), foot (5), face (4), brow (4), lip (4), book (4), story (4), fingertip (3), turn (3), hair (3), own (3), word (3), friend (3), ear (3), knee (3), father (3), mind (3), breath (3), bed (2), cheek (2), palm (2), skin (2), blood (2), year (2), work (2), neck (2), faith (2), dog (2), throat (2), mouth (2)
	adolescent	hand (46), eye (26), mother (25), head (21), knife (19), face (18), dad (16), shoulder (16), side (15), foot (13), father (13), arm (12), finger (12), throat (10), hair (9), body (9), lip (8), back (8), word (8), mask (7), breath (6), cheek (6), ear (6), pocket (6), friend (5), house (5), door (5), work (5), brain (5), skin (5), mam (4), heart (4), mind (4), voice (4), blood (4), life (4), wrist (4), parent (4), baby (4), wasteland (3)
	adult	hand (44), eye (34), face (23), head (22), mouth (17), lip (17), finger (13), arm (13), breath (12), son (12), back (10), clothe (7), desk (7), tongue (6), fist (6), voice (5), hair (5), cigarette (5), brow (5), mate (5), cheek (5), dog (5), jacket (5), bag (5), pocket (5), own (4), knee (4), chest (4), shoulder (4), room (4), collar (4), foot (4), word (4), glass (4), palm (3), handkerchief (3), fingertip (3), smoke (3), footstep (3), office (3)
	oldadult	home (1), dog (1)
young adult (12–15)	child	eye (224), hand (213), head (136), face (96), arm (90), dad (79), name (76), finger (73), mum (64), heart (48), friend (47), sister (44), lip (43), mouth (42), brother (42), cheek (41), shoulder (41), foot (40), mother (38), hair (37), mind (33), side (33), voice (33), breath (32), skin (31), life (27), fist (26), father (26), tongue (22), pocket (22), body (19), way (18), parent (16), palm (16), room (16), knife (16), thought (16), back (15), bed (15), tree (15)

	adolescent	eye (492), hand (398), head (292), face (175), arm (161), mother (160), mum (131), finger (119), lip (116), heart (101), friend (88), dad (84), side (84), voice (77), shoulder (77), name (75), foot (74), cheek (74), skin (70), mouth (63), breath (62), tongue (59), mind (58), hair (54), back (54), father (50), fist (48), chest (47), body (46), ear (45), palm (45), throat (37), sister (37), mam (35), room (32), dream (32), parent (31), life (30), pocket (28), knife (28)
	adult	eye (231), hand (221), head (136), arm (111), face (110), voice (74), finger (66), lip (65), mouth (40), breath (38), son (36), house (33), shoulder (30), side (29), hair (28), back (25), tongue (24), leg (23), cheek (22), foot (21), fist (20), pocket (19), way (19), skin (17), throat (17), chest (16), tooth (16), boy (16), body (16), knee (14), name (14), life (14), heart (13), door (13), baby (12), word (11), book (11), dog (11), palm (10), fire (10)
	oldadult	eye (91), hand (47), head (31), finger (25), face (20), arm (18), book (16), lip (13), voice (13), shoulder (11), bed (11), body (10), cheek (10), door (7), tale (7), one (7), throat (6), helmet (6), tea (6), foot (6), pocket (5), mouth (5), hair (5), breath (5), song (5), desk (5), helper (5), son (4), room (4), palm (4), window (4), back (4), jacket (4), Corinna (4), pencil (4), beard (4), Helper (4), suit (3), tongue (3), soldier (3)
	child	eye (119), hand (110), head (98), dad (75), mum (57), name (43), wing (42), face (41), mouth (40), arm (40), side (36), friend (35), shoulder (31), ear (31), voice (26), foot (25), hair (22), heart (21), lip (21), book (19), finger (18), mind (17), back (17), mam (16), chest (16), leg (16), tongue (13), table (13), pocket (12), life (12), uncle (11), lap (11), box (11), Mam (11), skin (10), time (10), family (10), way (10), room (10), cheek (9)
children (7–9)	adolescent	lass (1), lad (1), teacher (1), pal (1), behalf (1), backside (1), mouth (1), mummy (1), hand (1), eye (1), trembling (1), dad (1), family (1), workplace (1), paper (1)
	adult	hand (41), head (29), eye (24), arm (19), voice (16), face (15), finger (9), name (9), shoulder (8), brother (7), notebook (7), hair (6), way (5), foot (5), smile (5), salon (5), bag (5), van (5), fist (4), uniform (4), cheek (4), nephew (4), sister (4), apartment (4), son (3), leg (3), word (3), knee (3), work (3), mouth (3), lip (3), ambition (3), side (3), office (3), whistle (3), lap (3), colleague (3), house (2), footstep (2), laughter (2)

	oldadult	eye (12), hand (11), face (10), head (9), voice (8), name (8), brother (7), palm (5), nephew (4), sister (4), apartment (4), foot (3), lip (3), shoulder (3), ambition (3), caravan (2), stick (2), garden (2), lap (2), knee (2), love (2), finger (2), job (2), way (2), report (2), forefinger (1), fingertip (1), figure (1), prayer (1), limp (1), day (1), ancient (1), frame (1), mate (1), door (1), pyjama (1), stuff (1), brow (1), dad (1), mother (1)

3. List of 40 most common adjectives

intended readership	characters' age	terms
adult (18+)	child	good (7), little (7), lovely (4), old (3), other (3), small (3), new (2), brave (2), many (2), poor (2), young (2), invisible (1), slow (1), awful (1), bonny (1), bliddy (1), inner (1), scorched (1), wild (1), chesty (1), sacrificial (1), normal (1), endless (1), gorgeous (1), beautiful (1)
	adolescent	good (14), other (5), poor (4), new (4), young (3), little (3), lovely (3), bonny (3), sweet (2), silly (2), clever (1), many (1), old (1), tall (1), only (1), brainy (1), brutal (1), open (1), troubling (1), different (1), graceful (1), dozy (1), filthy (1), tawdry (1), stupid (1), own (1), true (1), lazy (1), temporary (1), canny (1), proper (1), prim (1)
	adult	poor (4), other (4), lovely (3), own (3), stupid (3), cold (2), dark (2), english (2), good (2), kind (1), true (1), strict (1), fat (1), fine (1), hatted (1), fuckin (1), clumsy (1), weird (1), normal (1), squat (1), muscular (1), little (1), shocked (1), visible (1), bloody (1), miserable (1), bonny (1), famous (1), desperate (1), wizened (1), gentle (1), young (1), aged (1), handsome (1), right (1), canny (1), sane (1), fellow (1), old (1)
	oldadult	old (2)
young adult (12–15)	child	little (113), young (23), new (18), good (15), bonny (15), poor (14), lovely (14), skinny (12), stupid (7), bloody (6), other (5), silly (4), bad (4), tiny (4), best (4), whole (3), haired (3), mine (3), ancient (3), ordinary (3), fifth (3), delicate (2), old (2), ignorant (2), scared (2), strong (2), own (2), lucky (2), blond (2), precious (2), ill (1), funny (1), great (1), invalid (1), right (1), runaway (1), legged (1), small (1), thin (1), hyena (1)
	adolescent	good (42), old (16), little (15), dead (13), aged (11), lovely (10), poor (10), stupid (9), bad (9), young (7), big (7), black (7), skinny (6), other (6), bonny

		(5), brave (5), strange (5), lucky (4), ordinary (4), troubled (4), weird (4), sweet (3), great (3), daft (3), bloody (3), clever (3), silly (3), ancient (3), white (3), hipster (2), beautiful (2), dreamy (2), single (2), own (2), bliddy (2), true (2), brilliant (2), different (2), shy (2), scrawny (2)
	adult	poor (19), old (19), lovely (14), young (11), stupid (8), little (7), tall (5), other (5), boring (5), daft (5), extraordinary (4), big (4), haired (4), nice (3), first (3), beautiful (3), crazy (3), foster (2), white (2), tasty (2), good (2), deep (2), silly (2), great (2), useless (2), top (2), straight (2), awful (2), reckless (2), drunken (2), dead (2), impatient (1), low (1), precious (1), troubled (1), favourite (1), stinking (1), strange (1), rotten (1), thin (1)
	oldadult	old (22), ancient (2), filthy (2), own (2), lovely (2), pale (1), frail (1), kind (1), trembling (1), strange (1), paradoxical (1), irish (1), quick (1), gentle (1), same (1), poor (1), bad (1), good (1), little (1), angry (1), tall (1)
children (7–9)	child	little (70), young (18), good (17), lovely (17), poor (12), other (9), german (8), new (8), silly (6), dear (5), ordinary (4), lonely (3), same (3), different (3), great (3), lost (3), clever (3), few (3), small (3), bright (2), stupid (2), natural (2), ickle (2), skinny (2), old (2), scrawny (2), brave (2), perfect (2), nice (2), fantastic (2), weird (2), cheeky (2), bloomin (2), gifted (2), excited (2), several (2), wingy (2), reet (2), headed (1), tiny (1)
	adolescent	few (1), other (1), certain (1), criminal (1)
	adult	great (10), lovely (9), young (9), new (7), burly (7), other (6), big (6), little (6), good (5), poor (5), next (4), haired (4), dark (4), weird (3), same (3), nice (3), identical (3), twin (3), silly (3), chubby (3), huge (3), barmy (2), slender (2), winged (2), sensible (2), clever (2), daft (2), old (2), homeless (2), wise (1), square (1), full (1), wobbly (1), close (1), ordinary (1), many (1), terrible (1), correct (1), sorry (1), absent (1)
	oldadult	old (12), young (6), identical (3), twin (3), sweet (2), nice (2), little (2), winged (2), daft (1), smelly (1), silly (1), boomy (1), stepped (1)

Appendix D: list of novels included in the corpus of chapter seven and the corresponding age of the intended reader

Author	Title (year of publication)	Minimal age of the intended reader
Bibi Dumon Tak	Mikis de ezeljongen (2011)	7
	Bezem (2005)	7
	Latino King (2010)	15
	Rotjongens (2007)	18 (book)
	De dag dat ik mijn naam veranderde (2020)	18 (book)
Ed Franck	Els heeft een pels (1995)	7
	Moet je echt weg? (1987)	7
	Spetters op de kermis (1985)	9
	Wie wil Wubbe weg? (1989)	9
	Begraaf me over de bergen (1988)	18
Guus Kuijer	Tin Toeval en de kunst van Madelief (1989)	7
	Ik ben Polleke, hoor! (2001)	9
	Met de wind mee naar zee (2001)	9
	Het land van de neushoornvogel (1985)	12
	Drie verschrikkelijke dagen (1976)	12
	Een gat in de grens (1975)	12
Bart Moeyaert	Die steeg van ons (1995)	7
	Mansoor (1996)	9
	Suzanne Dantine (1989)	12
	Terug naar af (1986)	12
	Broere (2000)	15
	Het is de liefde die we niet begrijpen (1999)	15
	Graz (2009)	18
Hilde Vandermeeren	Moord in de wijk (2019)	18 (book)
Joke van Leeuwen	Maar ik ben Frederik, zei Frederik (2013)	7
	Toen ik (2017)	7
	Ik heet Reinier en ons huis is afgebrand (2020)	9
	Slopie (2004)	9
	Toen mijn vader een struik werd (2010)	9
	Dit boek heet anders (1992)	9
	Het verhaal van Bobbel (1987)	9
	Iep! (1996)	9
	Kukel (1998)	9
	Wijd weg (1991)	9
	Bezoekjaren (1988)	12
	Alles nieuw (2008)	18 (website author)
	De onervarenen (2015)	18 (website author)
	De tjilpmachine (1990)	18 (book)
	Feest van het begin (2012)	18 (website author)
	Hier (2018)	18 (website author)
	Vrije vormen (2002)	18 (website author)

Appendix E: word lists of actions, possessions and adjectives associated with female and male older adult characters in children's literature (CL, ages 7–9), young adult literature (YA, ages 12–15) and adult literature (AL, ages 18+)

1. word list of actions

Topics		Female	Male
Movement	CL	stand, walk (no longer well), sit, come, pull, stop, shuffle	stand, walk (further), sit, put, step, fall, sneak, hide (well), come (close), lie, disappear, lift (high)
	YA	come, stand, walk, stop, sit, disappear, step, appear (finally), pull (loudly), trip (backwards), hoist (laboriously), ram (furiously), shuffle (softly), stoop (laboriously), bend (laboriously)	stand (encouragingly), shuffle, lie (dead still), walk, sit (motionless), lean, get lost, hand over, appear, stamp (frantically), come, turn around (quickly), move
	AL	come, stand (later), sit (naturally), walk, pull, step (carefully), push open (wider), shake (vigorously), lie (later), close (well)	come (further), stand, lie (unhappily), sit, walk (slowly, higher), slide (far, open), step, push, pull (close), throw, shift (laboriously), stroke (rhythmically), drag (laboriously), throw (hard)
Perception	CL	see, know, forget, think, look at, learn, keep (spectating), hear, find (regretful, heavy, cosy)	know, think, see (well), read (amazingly fast), hear, listen, find (enjoyable, boring)
	YA	see (well), know, think, hear, understand, view (inquisitively), watch (pervasively), feel	know, see, remember (nicely), look out, hear, think
	AL	think (suddenly), know (precisely, probably, better), see (suddenly, less sharp, less well), hear, know (all along), read (daily), remember (well), believe, forget, hope (fervently), look (close), find (regretful, fine, useful, pretty)	see, think, know (for sure), hear, reflect (deeper), understand (well), read (on), find (funny, pleasantly empty, rude, pretty)
Expression and emotion	CL	say, talk, tell, ask, become (cheerful), mean, point, forbid (strictly), look (angry, grateful)	say (always, sadly, gruffly, sternly, conspiratorially), tell (on), ask, write, talk (very softly, too loud), sigh (deeply), laugh (bitterly), smile (wide), whisper (incoherently, agitated), recite, punish, hit, comfort, look (sad, dreamy, encouragingly, friendly)
	YA	say (thoughtfully, urgently, nicely, quickly, almost unintelligibly), ask, whisper, answer (quickly), convince, shout, nod, cry, tell, laugh (nervously), hit, giggle, wail (passionately), point, look (anxious)	say (despondently, clear), tell (on), shut up, nod (bitterly, friendly), ask, shout, smile, giggle, hit, answer, speak, chuckle, mumble, repeat (literally), write, call, laugh, look (elated, serene)
	AL	say (softly, angrily, cheerfully), ask (differently), tell (imperturbably continue, long-windedly, on), write (secretly), call (desperately), shout (anxiously, helplessly), nod, speak (very articulately), greet (briefly), sound (angry), lie (well), talk (incessantly), feel (lonely, miserable,	say (directly, conspiratorially), ask, tell, shout, hit (hard), call, speak, cry, repeat, interfere, become (mad, grumpy), look (suspicious)

		restless, good), look (satisfied, serious, good, weird), become (uncomfortable)	
Illness and older age	CL	live, die, breathe (slowly, ever more slowly), exist	care (badly), feel (weak), die, be (too old and too skinny), live (for so long already)
	YA	breathe (excessively), lie (dead)	die, live, leave behind
	AL	die, feel (sick), become (too old, old), live (too long), hang (on), look (sick), faint	feel (ancient), live (even longer), pant (rather quick these days)
Leisure activities	CL	bake, sleep, shoot, draw (well)	help, draw (well), clean
	YA	wake (suddenly), work, drink	plant (deep), work, build, chew, drive (further), knead
	AL	wake, sleep, polish (shinily), scrub (too hard), kiss (royally)	re-enact (later)

2. word list of possessions

Topics		Female	Male
Body parts	CL	arm, face, eye, leg, vein, hand, head (small), mouth, hair (long), eyebrow, haircut, voice, skin, lap, ear, breath, muscle, bone	hand (flat, skinny), head, face, hair, ear, nose, arm, eyebrow, lung, eye, body (skinny), back, mouth (old), breath, voice, finger (bony, old), corner of the mouth (old), skin (unwashed), knee, lip, eyelid, forehead, tongue (old), cheek, leg, wrinkle
	YA	hand (dry, bony), head, eye, nose, arm (fleshy), finger (chubby), leg, mouth, lip, foot, voice, body, breath (last), shoulder, look, tongue, bone, face, knee, back, cheek, ear, thumb, index finger, chest	eye, face (greyish-brown), arm (other), spirit, hand (cardboard, bony), heart, body (wiry), hair (thin), leg, neck, chin, earlobe, tooth, skull, right hand, heel, mouth, look, voice, lip, knee, beady eye, belly, rib, nose (sharp), head, shoulder, ear, finger, joint, tongue, lap, leg
	AL	head, eye (dry), hand, body (worn-out, rigid), hair (grey, tangled, thin white, unbrushed), arm, skin, back (uncooperative, stiff), chest, shoulder, finger (old), cheek, face, mouth, heart, leg, foot (bare), blood, voice, forehead, lap, belly, skeleton, nose (pale)	hand (blotchy), leg (frail, weak), eye, arm, head, face (fat), back (hairy, hunched), body (heavy, unwieldy, slow), breath, finger, belly, index finger, knee, heart, look, shoulder, mouth, voice, cheek (cold), foot, neck (thick), forehead, whites of the eyes (yellowish)
Health	CL	life, glasses	dentures, glasses, health, life (entire), existence, grave
	YA	death, deathbed	movement, wheelchair
	AL	glasses (thick), death, deathbed	death, birthday (own)
Community	CL	neighbour, locksmith, grandmother, child, student, class, grandchild, man	brother, friend (new), father, aunt, wife, grandfather, mother
	YA	man, daughter (fat), marriage, friend	grandson, daughter, grandfather, great-grandfather, grandchild, people (former), customer (first)
	AL	daughter (only-begotten, absent), man, grandchild, parent, mother, family, sister (older)	daughter, child (dearest), grandchild, murderer, granddaughter (youngest), son, wife (dead)
Home	CL	bed, bedside table, pillow, desk, blanket	house (new), scooter, garden, chair, study (large), desk

	YA	bedside table, room	attic window, carpet, window, house (pretty), chair, tower, bed (warm)
	AL	living room, chair (rattan), house, bed, kitchen, letterbox	bed (own), chair (lazy), house (own), room, footboard
Clothing	CL	coat, clothes	shoelace, uniform (old), pocket, clothes (dusty)
	YA	apron (grubby), boot, glove, nurse's uniform (old), sleeve, coat, rain cap, slipper	shirt, hat, beanie, clothes, coat
	AL	skirt, dress (best)	coat, shoe, jacket, pocket, trousers, shirt, underpants (white), sock (black), customs clothes (pretty)

3. word list of adjectives

Topics		Female	Male
Appearance	CL	grey	fat, red, skinny, dusty, small
	YA	small, grey, wrinkly	skinny, white (before), unwashed, thin
	AL	flat, blotchy (already a bit less), tall, sweaty	tall, heavy (too)
Emotions	CL		sad, scared (not), angry (maybe, very, also), happy (maybe)
	YA	happy	dissatisfied
	AL	happy (always, not), scared, desperate (again), grateful, upset (very), at ease (not), relieved (very)	happy, angry, proud
Illness	CL	sick (very)	tired, sick
	YA		tired
	AL	tired	
Character traits	CL	patient, interesting (perhaps more), thrifty	kind, old-fashioned, strict, well-behaved, clever, valuable
	YA	tough, normal, kind (always), severe, honest	secretive
	AL	curious (very)	wise, climate neutral (not), interesting, good
Age	CL	old (themselves, rather, very)	old (even more, so)
	YA	old	old, ancient
	AL	old, younger (maybe)	old, birthday-y

List of tables

List of figures

Author contributions

Research project	
Funding acquisition	Vanessa Joosen
Chapter 1	
Writing	Lindsey Geybels
Chapter 2	
Conceptualisation	Lindsey Geybels, Wouter Haverals and Vanessa Joosen
Programming	Lindsey Geybels and Wouter Haverals
Analysis and interpretation	Lindsey Geybels, Wouter Haverals and Vanessa Joosen
Writing	
- Introduction	Lindsey Geybels, Wouter Haverals and Vanessa Joosen
- Research into Crosswriters	Lindsey Geybels
- Stylometric method	Wouter Haverals
- Case study Hilde Vandermeeren	Wouter Haverals and Vanessa Joosen
- Case study Guus Kuijer	Wouter Haverals and Vanessa Joosen
- Case study English titles	Lindsey Geybels
- Case study J.K. Rowling	Lindsey Geybels
- Conclusion	Lindsey Geybels, Wouter Haverals and Vanessa Joosen
Revising and editing	Lindsey Geybels, Wouter Haverals and Vanessa Joosen
Chapter 3	
Conceptualisation	Lindsey Geybels and Wouter Haverals
Programming	Lindsey Geybels, Wouter Haverals and Mike Kestemont
Analysis and interpretation	Lindsey Geybels and Wouter Haverals
Writing	
- Introduction	Lindsey Geybels and Wouter Haverals
- Diverging schemes	Lindsey Geybels
- Digital analyses	Lindsey Geybels and Wouter Haverals
- Form in the series	Wouter Haverals
- From childlike to mature topics	Lindsey Geybels
- Conclusion	Lindsey Geybels and Wouter Haverals
Revising and editing	Lindsey Geybels and Wouter Haverals
Chapter 4	
Conceptualisation	Lindsey Geybels

Programming	Lindsey Geybels
Analysis and interpretation	Lindsey Geybels
Writing	Lindsey Geybels
Revising and editing	Lindsey Geybels

Chapter 5	
Conceptualisation	Lindsey Geybels
Programming	Lindsey Geybels
Analysis and interpretation	Lindsey Geybels
Writing	Lindsey Geybels
Revising and editing	Lindsey Geybels

Chapter 6	
Conceptualisation	Lindsey Geybels and Vanessa Joosen
Programming	Lindsey Geybels and Pieter Fivez
Analysis and interpretation	Lindsey Geybels
Writing	Lindsey Geybels
Revising and editing	Lindsey Geybels and Vanessa Joosen

Chapter 7	
Conceptualisation	Lindsey Geybels
Programming	Lindsey Geybels and Pieter Fivez
Analysis and interpretation	Lindsey Geybels
Writing	Lindsey Geybels
Revising and editing	Lindsey Geybels

Chapter 8	
Writing	Lindsey Geybels

Summaries

English summary

Fiction for young readers serves as a powerful tool for shaping young minds and conveying ideological messages. Scholars in the field of children's literature studies have extensively explored the ideological load carried by literature for children, including in how age is constructed and represented. The underlying assumption in previous research is that ideology is inherent in every form of discourse and utterance. The age ideologies, or age norms, embedded in fiction influence the age socialisation process and shape children's understanding of age within a broader social context. They are present in both implicit and explicit ways, which can aid children in learning to interact with others but also perpetuate harmful biases. However, age norms are not only present in books for children, but also in young adult literature and adult novels and have just as much potential to influence readers from other stages in life. Therefore, a critical examination of fiction for younger readers and adult literature provides valuable insights into the age ideologies at play. This study focuses on the concept of the intended reader, examining both the sociological aspect of the reader's age and the literary representations of age in fiction from a discursive standpoint. This dissertation shows that the construction of age in fiction varies according to the age of the intended reader in different ways.

The first research objective of this dissertation is to gain deeper understanding of the construction of age for young readers, and this on two levels; surface features of the text and the representation of fictional age. To answer the research questions tied to this aim, this dissertation makes use of an innovative methodology, made up of methods for digital text analysis. Compiling a set of methods that can be used to study age-related questions in fiction for young readers is the second research objective of this dissertation and addresses a significant research gap by expanding the scale and scope of previous studies conducted in this area. By combining different approaches, the chapters in this dissertation contributes to a more comprehensive understanding of age-related themes in children's literature, young adult literature, and fiction for adults.

To investigate the influence of the age of the intended reader on the construction of age by a work of fiction, or how the characteristics of a text are affected by the age markers placed on books, surface features are examined in chapters two through six. The methods that are used to study this level of the construction of age are stylometric analyses, readability measures, topic modelling, statistical models, and a general comparison of surface features. In chapter two, stylometric tools are applied to the works of ten Dutch and English crosswriters to identify individual stylistic differences and similarities among the authors. The analysis takes into account the age of the intended reader as well as the publication date to examine their influence on writing style. Several case studies delve into specific authors and a corpus-wide of English titles to identify general trends and outliers. The findings suggest that writing style is more closely aligned with the age of the intended reader than the time period of writing, with young adult literature being stylistically closer to adult literature than books for younger children. The style associated with younger readers is distinct and transcends individual author differences.

Considering the valuable contributions of this study, the stylometric analyses in this chapter are revisited in chapter six. While the study acknowledges certain limitations in terms of paratextual features and lacks a discussion on the effect of the age of the intended reader on the entire corpus, these gaps are addressed in subsequent chapters. The stylometric analyses in chapter six focus on the oeuvre of British author David Almond, which also show stylistic clustering according to the age of the intended reader.

The analyses in chapter three emerge from the lack of attention the social construction of age has received in children's literature studies compared to research on gender, race, and class. This chapter focuses on the construction of age in J.K. Rowling's *Harry Potter* series in relation to the intended reader's age, which is renamed 'implied reader' in this chapter for reasons of editorial preference. The study examines the correlation between the evolution of form and content in the series and the age of its implied readership. By scrutinising existing guidelines and conducting digital analyses on style and topics, the article contributes to discussions on readability measures, the inclusion of explicit age markers, and the evolution of features in the *Harry Potter* books. The chapter notes the importance of supplementing quantitative analyses with close reading and provides insights into the terms 'complexity' and 'simplicity' commonly used when discussing differences between children's and adult fiction. These terms are further elucidated in chapter four, which applies statistical analyses to the entire corpus of the dissertation, totalling 738 texts. The analyses highlight the potential of computational techniques to convert words into numerical data, allowing for the examination of certain aspects on a large scale. Elements like sentence length, word length, and lexical diversity, associated with writing style and text complexity, are statistically evaluated to determine their corelation with the age of the intended reader. The results demonstrate that these textual elements are more effective in detecting the age of the intended reader than the author's identity or age, aligning with findings from previous chapters.

Chapter five further contributes to the discussion on differences between literature for children and adults by using digital tools to examine the works of crosswriter Joke van Leeuwen, who writes for both readerships. The study deconstructs readability formulas, as previously done in chapters three and four, and identifies formal differences in word and sentence length, lexical diversity, and the ratio of dialogue to narrative. Additionally, chapter five acts as a bridge between the two research questions tied to the first research objective, exploring the age constructed by the text and studied through formal features, as well as the age portrayed within the text through character distribution and explicit statements. The chapter provides insights through metareflections and contextual analysis, distinguishing it from the other chapters in the dissertation. The analyses in this chapter incorporate both raw text files and annotated files, offering a closer examination of the texts as well as the author's views on writing for different age groups. This methodology is carried through to chapter six, which adds analyses on characterisation to complement the stylometric tools applied to the oeuvre of David Almond as elaborated on above. By employing topic modelling on character speech and syntactic parsing, this chapter examines the ideologies surrounding fictional age as depicted in Almond's characters. While the different types of analyses may result in a less cohesive conclusion, they provide a comprehensive picture of the construction of age in Almond's oeuvre.

Chapter seven expands the use of the syntactic parser introduced in chapter six and addresses the intersection of age and gender representation in fiction, with a focus on older adult characters, through the analysis of 41 Dutch books intended for readers of different ages. The

findings reveal that older adult characters are more prevalent in fiction for adults, where they are often depicted as vulnerable to illness, experiencing physical decline, and having limited social networks. In contrast, children's books show little association between older adulthood and mortality. Ageist stereotypes are present throughout the corpus, and male older characters are more associated with physicality, while female older characters exhibit a broader range of character traits and emotions. The chapter extends previous use of the syntactic parser by considering the intersection of age and gender as well as taking into account modifier of the verbs, adjectives, and grammatical possessions that are examined.

Nederlandstalige samenvatting

Fictie voor jonge lezers draagt bij tot de ontwikkeling van jonge geesten door ideologische boodschappen voort te dragen. Onderzoekers van jeugdliteratuur hebben reeds uitgebreid onderzoek gedaan naar de ideologische lading van literatuur voor kinderen, onder andere naar de manier waarop leeftijd wordt geconstrueerd en gerepresenteerd. De onderliggende aanname in eerder onderzoek is dat ideologie inherent is aan elke vorm van discours en uiting. De ideologieën omtrent leeftijd, of leeftijdsnormen, die zijn ingebed in fictie beïnvloeden het socialisatieproces en geven voor kinderen vorm aan het begrip van leeftijd binnen een bredere sociale context. Ze zijn zowel impliciet als expliciet aanwezig, wat kinderen kan helpen bij het leren omgaan met anderen, maar ook schadelijke vooroordelen in stand kan houden. Leeftijdsnormen komen echter niet alleen voor in boeken voor kinderen, maar ook in literatuur voor jongvolwassenen en romans voor volwassenen en hebben daar net zoveel potentieel om lezers te beïnvloeden. Daarom biedt een kritisch onderzoek naar fictie voor jongere lezers en literatuur voor volwassenen waardevolle inzichten in de leeftijdsideologieën die hierbij een rol spelen. Deze studie richt zich op het concept van de beoogde lezer, waarbij zowel het sociologische aspect van de leeftijd van de lezer als de literaire representaties van leeftijd in fictie vanuit een discursief standpunt worden onderzocht. Dit proefschrift laat zien dat de constructie van leeftijd in fictie op verschillende manieren varieert afhankelijk van de leeftijd van de beoogde lezer.

Het eerste onderzoeksdoel van dit proefschrift is om meer inzicht te krijgen in de constructie van leeftijd voor jonge lezers, en wel op twee niveaus: in de oppervlaktekenmerken van de tekst en in de representatie van fictionele leeftijd. Om de onderzoeksvragen te beantwoorden die aan dit doel gekoppeld zijn, maakt dit proefschrift gebruik van een innovatieve methodologie, bestaande uit methoden voor digitale tekstanalyse. Het samenstellen van een set methoden die gebruikt kan worden om leeftijdsgerelateerde vragen in fictie voor jonge lezers te bestuderen is het tweede onderzoeksdoel van dit proefschrift en vult een belangrijke onderzoeksleemte door de omvang en reikwijdte van eerdere studies op dit gebied uit te breiden. Door verschillende benaderingen te combineren, dragen de hoofdstukken in dit proefschrift bij aan een uitgebreider begrip van leeftijdsgerelateerde thema's in jeugdliteratuur, adolescentenliteratuur en fictie voor volwassenen.

Om de invloed van de leeftijd van de beoogde lezer op de kenmerken van een tekst te onderzoeken, worden in de hoofdstukken twee tot en met zes oppervlaktekenmerken onderzocht. De methoden die gebruikt worden om dit niveau van leeftijdsconstructie te bestuderen zijn stylometrische analyses, leesbaarheidsmetingen, 'topic modelling', statistische modellen en een algemene vergelijking van oppervlaktekenmerken. In hoofdstuk twee worden stylometrische methodes toegepast op het werk van tien Nederlandse en Engelse multipublieksauteurs om individuele stijlverschillen en overeenkomsten tussen de auteurs te identificeren. De analyse houdt rekening met de leeftijd van de beoogde lezer en de publicatiedatum om hun invloed op de schrijfstijl te onderzoeken. Verschillende casestudies gaan dieper in op specifieke auteurs alsook een alle Engelstalige titels in het corpus om algemene trends en uitschieters te identificeren. De bevindingen suggereren dat de schrijfstijl nauwer samenhangt met de leeftijd van de beoogde lezer dan met de periode waarin het boek is geschreven, waarbij literatuur voor jongvolwassenen stilistisch dichter bij literatuur voor volwassenen staat dan bij boeken voor jongere kinderen. De stijl die geassocieerd wordt met jongere lezers is verschillend en overstijgt individuele auteursverschillen. Gezien de waardevolle bijdragen van dit onderzoek, komen de

stylometrische analyses in dit hoofdstuk terug in hoofdstuk zes. Hoewel de studie bepaalde beperkingen erkent op het gebied van paratekstuele kenmerken en het effect van de leeftijd van de beoogde lezer op het hele corpus niet bespreekt, worden deze hiaten in volgende hoofdstukken behandeld. De stylometrische analyses in hoofdstuk zes richten zich op het oeuvre van de Britse auteur David Almond, dat eveneens stilistische clustering vertoont op basis van de leeftijd van de beoogde lezer.

De analyses in hoofdstuk drie komen voort uit het gebrek aan aandacht dat de sociale constructie van leeftijd heeft gekregen in onderzoek naar jeugdliteratuur in vergelijking met onderzoek naar gender, ras en klasse. Dit hoofdstuk richt zich op de constructie van leeftijd in de *Harry Potter* serie van J.K. Rowling in relatie tot de leeftijd van de beoogde lezer, die in dit hoofdstuk om redactionele redenen 'impliciete lezer' wordt genoemd. De studie onderzoekt de correlatie tussen de evolutie van vorm en inhoud in de serie en de leeftijd van zijn beoogde lezerspubliek. Door bestaande richtlijnen onder de loep te nemen en digitale analyses van stijl en onderwerpen uit te voeren, draagt het artikel bij aan discussies over leesbaarheidsmetingen, het opnemen van expliciete leeftijdsaanduidingen op boeken en de evolutie van kenmerken in de *Harry Potter*-boeken. Het hoofdstuk wijst op het belang van het aanvullen van kwantitatieve analyses met een gedetailleerdere lezing en geeft inzicht in de termen 'complexiteit' en 'eenvoud' die vaak worden gebruikt bij het bespreken van verschillen tussen kinder- en volwassenenfictie. Deze termen worden verder toegelicht in hoofdstuk vier, waarin statistische analyses worden toegepast op het gehele corpus van het proefschrift, in totaal 738 teksten. De analyses benadrukken het potentieel van computationele technieken om woorden om te zetten in numerieke gegevens, waardoor bepaalde aspecten op grote schaal kunnen worden onderzocht. Elementen zoals zinslengte, woordlengte en lexicale diversiteit, die geassocieerd worden met schrijfstijl en tekstcomplexiteit, worden statistisch geëvalueerd om hun correlatie met de leeftijd van de beoogde lezer te bepalen. De resultaten tonen aan dat deze tekstuele elementen effectiever zijn in het detecteren van de leeftijd van de beoogde lezer dan de identiteit of leeftijd van de auteur, wat overeenkomt met bevindingen uit eerdere hoofdstukken.

Hoofdstuk vijf draagt verder bij aan de discussie over verschillen tussen literatuur voor kinderen en volwassenen door digitale hulpmiddelen te gebruiken om het werk van multipublieksauteur Joke van Leeuwen te onderzoeken, die voor beide lezersgroepen schrijft, alsook voor kinderen van verschillende leeftijden. De studie deconstrueert leesbaarheidsformules, zoals eerder gedemonstreerd in hoofdstuk drie en vier, en identificeert formele verschillen in woord- en zinslengte, lexicale diversiteit en de verhouding tussen dialoog en verhaal. Daarnaast fungeert hoofdstuk vijf als een brug tussen de twee onderzoeksvragen die verbonden zijn aan het eerste onderzoeksdoel, door de leeftijd te onderzoeken die door de tekst wordt geconstrueerd en bestudeerd op vlak van formele kenmerken, evenals de leeftijd die in de tekst wordt uitgebeeld in personageverdeling en expliciete uitspraken. Het hoofdstuk biedt inzichten door middel van metareflecties en contextuele analyse, waardoor het zich onderscheidt van de andere hoofdstukken in het proefschrift. De analyses in dit hoofdstuk bevatten zowel de basis tekstbestanden als geannoteerde teksten, waardoor de werken en de opvattingen van de auteur over het schrijven voor verschillende leeftijdsgroepen nader worden onderzocht. Deze methodologie wordt voortgezet in hoofdstuk zes, waarin analyses van karakterisering worden toegevoegd als aanvulling op de stylometrische hulpmiddelen die zijn toegepast op het oeuvre van David Almond, zoals hierboven beschreven. Door gebruik te maken van 'topic modelling' op karakterspraak en syntactische parsing, onderzoekt dit hoofdstuk de ideologieën rond fictieve leeftijd

zoals weergegeven in Almonds personages. Hoewel de verschillende soorten analyses leiden tot een minder samenhangende conclusie, geven ze een uitgebreid beeld van de constructie van leeftijd in het oeuvre van Almond.

Hoofdstuk zeven breidt het gebruik van de syntactische parser geïntroduceerd in hoofdstuk zes uit en gaat in op het snijvlak van leeftijd en genderrepresentatie in fictie, met een focus op oudere volwassen personages, door middel van de analyse van 41 Nederlandse boeken bedoeld voor lezers van verschillende leeftijden. De bevindingen laten zien dat oudere volwassen personages vaker voorkomen in fictie voor volwassenen, waar ze vaak worden afgeschilderd als kwetsbaar voor ziekte, lichamelijk achteruitgaan en beperkte sociale netwerken hebben. In kinderboeken is er daarentegen weinig verband tussen oudere volwassenheid en sterfelijkheid. Leeftijds-gebonden stereotypen komen in het hele corpus voor en mannelijke oudere personages worden meer geassocieerd met lichamelijkheid, terwijl vrouwelijke oudere personages een breder scala aan karaktereigenschappen en emoties vertonen. Het hoofdstuk breidt het eerdere gebruik van de syntactische parser uit door rekening te houden met het snijpunt van leeftijd en geslacht en met de bijwoorden van de onderzochte werkwoorden, bijvoeglijke naamwoorden en grammaticale bezittingen.

Acknowledgments

This dissertation is part of the research project 'Constructing Age for Young Readers'. This project was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 804920).

Reading the Reader from a Distance would not have been possible without the support of the community of wonderful people with whom I am privileged to be surrounded, both at home and at the University of Antwerp. Firstly, I express my heartfelt gratitude to my husband, Bart, who wholeheartedly supported my decision to embark on this academic journey and who is an exceptional father to our tiny human toddler. I am immensely grateful to my grandparents and in-laws, as they have been unwavering in their support, standing by my side during challenging times and uplifting me when I felt discouraged. The academic world can be demanding, but I am fortunate to say that I have not personally experienced its hardships, thanks to my colleagues, especially the members of our children's literature team. They are an exceptional group, providing not only unwavering support but also invaluable insights into my own growth throughout the years. There are numerous individuals I would like to thank, but too many moments to bring up. My gratitude extends to my supervisors, Vanessa and Mike, who presented me with an incredible opportunity, believed in me long before I did, and consistently provided encouragement. I would also like to acknowledge my CAFYR colleagues: Leander, with whom I had video-lunches during lockdown to maintain our sanity; Wouter, who generously shared his expertise in Python programming; Emma-Louise and Frauke, who serve as my role models as loving and successful mom-bosses; and Michelle, who kept my office chair warm and cosy while I was on leave. My newest office-mates, Maureen and Lien, with their infectious enthusiasm and imaginary pompons, have also played a significant role in shaping my journey thus far.

Additionally, I extend my gratitude to everyone who assisted me in various aspects of my research. I am thankful to the interns, volunteers, and colleagues who proofread texts, as well as the students, interns, and colleagues who diligently annotated texts. I would like to acknowledge the University of Antwerp library for acquiring several books that were essential additions to our corpus. Moreover, I express my appreciation to Mike Kestemont, Wouter Haverals, Pieter Fivez, and Jef Peeraer for their invaluable assistance with programming tasks. I received assistance from many more individuals throughout the different chapters of this dissertation, and I acknowledge their contributions as indicated in their respective publications.

Chapter 2: The authors would like to thank Mike Kestemont for providing feedback on the code used for the stylometric analyses in this article. A large part of the texts of the Dutch authors analysed in this article were provided by the publishing house Querido and the authors themselves. We are grateful for their contribution to our research. Because not all texts that make up the corpus of this article could be acquired in this way, a number of books were scanned, digitised and corrected. We thank Zoë Sierens and Ceylan Kuluçlu, two of the interns working on the CAFYR research project, for their precise work. We are grateful to Emma-Louise Silva for proofreading our work.

Chapter 4: The author would like to thank Vanessa Joosen, Mike Kestemont and Wouter Haverals 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.

Chapter 6: Thanks go out to the previous members of the CAFYR project, Wouter Haverals, Lien Goris and Lore De Greve, for their contribution in the early stages. The other members of the children's literature team at the University of Antwerp deserve a special note of thanks for acting as our sparring partners and contributing to the supportive and warm atmosphere in our research group: Andrea Davidson, Maureen Hosay, Paavo Van der Eecken, Miriam Wray and Yu Peng. Several students who participated in the Master's course 'Constructing Age for Young Readers' at the University of Antwerp contributed actively to the research project, by annotating books for our digital analyses. Several of them deserve special mention here: Miet Baelen, Lien Claeys, Jo De Bruyn, Anne Rosemarie Knobloch, Zoë Sierens, Olga van der Burgh, Merel Vanegeren and Loren Verreyen. Various volunteers from outside of the university also contributed to our project by helping to digitise books. Finally, David Almond has not only written the books that inspired this research, but he has also been exceptionally generous in making material available and answering our questions. The CAFYR research group is immensely grateful for this support and the intergenerational dialogues on various levels that he facilitated.

Chapter 7: The author would like to thank her colleagues at the children's literature research group at the University of Antwerp. In particular, the author acknowledges the invaluable support of postdoctoral researcher Emma-Louise Silva and expresses her gratitude for the lunchtime brainstorm session and the compelling hook. Further thanks goes out to Lien Claeys, a promising voluntary intern who painstakingly corrected data, CAFYR project leader Vanessa Joosen, for the quick an thorough feedback on the first draft of this article, and Leander Duthoy and Maureen Hosay for the encouraging words.