A style for every age: A stylometric inquiry into crosswriters for children, adolescents and adults

Wouter Haverals, Lindsey Geybels, and Vanessa Joosen
University of Antwerp, Antwerp, Belgium

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.

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, 1994; Stephens, 1999; Talley, 2011). 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, 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 (1991) 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, 2019: 65). Consequently, 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, 2020: 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’
ouevres. 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 recom mendations are usually meant as a minimal
age, since children’s books can have multiple addressees, including adults (Shavit, 1999).

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. Research into crosswriters

Several scholars, including Zohar Shavit (1999), Sandra Beckett (2009) and Helma Van Lierop-
Debrauwer (2000) 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, 2000: 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, 2009). 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 (2000) and Nicole
Hurkmans (2008) 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 (2013), 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 Janicke et al., 2015; Underwood, 2019).

Distant reading methods have only rarely been applied to crosswriters, but the scarce studies that do exist demonstrate their potential. Hurkmans (2008), 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 (2018) 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.

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, 1998). 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, 2001: 422), there is a great deal of evidence for their status as discriminators for an author’s writing style (Argamon and Levitan, 2005; Kestemont, 2014).
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, 2003). 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 (2013), 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, 2000). Some stylistic features might have been deliberately adopted, while others might have evolved rather unconsciously (Brandwood, 1992). 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 (1996) 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 (2014). By studying the linguistic development in the works of Samuel Beckett, they argue for a more nuanced periodisation of Beckett’s oeuvre. Interesting, 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 (Sikorska, 2013; Rudvin, 1994; Alvstad, 2010; Cermákova, 2018; Malmkjær, 2018; Toolan, 2018). 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 Franck, Kuijer, Moeyaert, Vandermeeren and Van Leeuwen. For English, we selected Almond, Fine, Gaiman, Pullman and 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., 2013). 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 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+).
As highlighted above, much of today’s stylometric research is based on the statistical occurrence of highly frequent words (Burrows, 2002; Binongo, 2003). For our analyses, we made use of the 100 most frequent words (MFW). These numbers are consistent with those used in similar stylometric research (Stamatatos, 2009; Van Hulle and Kestemont, 2016). 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, 1987). 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. 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 Cluster Analysis. 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, 1999). 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, 2013). 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. 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.

4. Case studies

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, 2002: 14). 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, 2001). Her own children have also commented on her drafts (AFT, 2007). 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). The hierarchical cluster analysis (Figure 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) 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, 2019: 10). 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.

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 gemenachte kind [The despised child], Kuijer (1980/1987: 125-136) depletes 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. Nevertheless, a PCA (based on the 100 MFW and for samples of 5000 words), shown in Figure 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. In the hierarchical cluster analysis (Figure 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, 1980: 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.).7 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 [1975, Throwing dolls] came into being’ (ibid.). Kuijer explains. Remarkably, both elements of this poetic statement are apparent from the PCA (Figure 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 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 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.

The first principal component (PC1) in Figure 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, 2009). In order to probe for deliberate choices, we carry out an additional Zeta test for contrastive text analysis (Burrows, 2007). 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 5. Several interesting observations can be made from this chart. What certainly stands out are the verb tenses (blue 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 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.

### 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. 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. The PCA is again generated using 100 MFW.

Figure 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 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 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+. When looking at the cluster analysis (Figure 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 (2017) wrote this book ‘for struggling, reluctant and dyslexic readers’ 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.

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 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’ (Lister, 2002). 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’ (2009). 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, 2009: 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.

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.

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 (Juola, 2013). The analyses presented in Figure 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, 2020).

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 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 9 shows the PCA for the unsegmented 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 8 and 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.

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.

Notes

1. 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 1999, xi-xii).

2. 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.

3. 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., 2015). 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).


5. 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.

6. My 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!’”

7. My translation. Original text: ‘Het dochtertje van de wasvrouw is een poging om […] een voudiger te gaan schrijven. [H]et staat natuurlijk interessanter als je ingewikkelder, komplexer, moeilijker schrijft […]. [M]aar als je voor kinderen gaat schrijven, dan kun je dat niet maken, die prikken daar gewoon doorheen’.

8. Studying both Dutch and English authors in the same analysis requires a time-consuming preprocessing, the method of which lies beyond the scope of this article (see Cinkova and Rybicki, 2020).

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