The nature and frequency of relative clauses in the language children hear and the language children read

Hsiao, Yaling; Dawson, Nicola J.; Banerji, Nilanjana; Nation, Kate

DOI:
10.1017/S0305000921000957

Preprint of article accepted for publication in Journal of Child Language. This article may not exactly replicate the final version published in the journal

The nature and frequency of relative clauses in the language children hear and the language children read: A developmental cross-corpus analysis of English complex grammar

Yaling Hsiao[1], Nicola J. Dawson[1], Nilanjana Banerji[2] & Kate Nation[1]

[1] University of Oxford

[2] Oxford University Press

Author note

The Oxford Children's Corpus is a growing database of writing for and by children developed and maintained by Oxford University Press for the purpose of children's language research. The work for this paper was supported by the British Academy Post-Doctoral Fellowship (PF2/180013) and the John Fell COVID Rebuilding Research Momentum Fund (0010144 CRRMF) awarded to Yaling Hsiao, a grant from the Nuffield Foundation (EDO/43392) to Kate Nation, and resources made available to Nilanjana Banerji by the department of Children's Dictionaries and Children's Language Data at Oxford University Press. Data and code associated with this paper are available on the Open Science Framework website (http://osf.io/p9kgr/). We thank Songjun He for research assistance. Correspondence concerning this article should be addressed to Yaling Hsiao, Department of Experimental Psychology, Anna Watts Building, University of Oxford, Oxford, OX2 6GG. E-mail: yaling.hsiao@psy.ox.ac.uk

Abstract

As written language contains more complex syntax than spoken language, exposure to written language provides opportunities for children to experience language input different from everyday speech. We investigated the distribution and nature of relative clauses in three large developmental corpora: one of child-directed speech (targeted at pre-schoolers) and two of text written for children, namely picture books targeted at pre-schoolers for shared reading and children's own reading books. Relative clauses were more common in both types of book language. Within text, relative clause usage increased with intended age, and were more frequent in nonfiction than fiction. The types of relative clause structures in text co-occurred with specific lexical properties, such as noun animacy and pronoun use. Book language provides unique access to grammar not easily encountered in speech. This has implications for the distributional lexical-syntactic features and associated discourse functions that children experience and from this, consequences for language development.

Keywords: grammatical development, reading, child-directed speech, corpus analysis, relative clauses, sentence processing

Word count: 10578

The nature and frequency of relative clauses in the language children hear and the language children read: a developmental cross-corpus analysis of English complex grammar

We do not write as we speak. Written language needs to represent meaning beyond the situation of the here and now, unaided by gesture, tone of voice and facial expression. To achieve its communication goals, written language has evolved to be more lexically diverse than spoken language; it also contains a higher proportion of complex and low-frequency syntactic structures (e.g., Biber, 1988; Roland, Dick, & Elman, 2007). Once children can read, they encounter language radically different from their day-to-day conversational experience. Our focus in this paper is with the nature of complex grammar children experience via written language. We investigated the frequency and use of different types of relative clause in three different corpora, one containing child-directed speech and two containing 'book language' – child-directed text written for children to read or to listen to in the context of shared reading. This allowed us to capture how children's linguistic input varies across spoken and written registers. We addressed how exposure to complex grammar varies developmentally, as experience with written language builds over time.

Relative clauses contain long-distance dependency relationships between their constituent elements that modify noun phrases, as shown in Table 1. It is well established that written language is generally more grammatically complex than speech. It involves more subordination and complementation (e.g., Biber, 1988; Halliday, 1989) and in a detailed linguistic analysis of adult corpora, Roland et al. (2007) found that the overall frequency of relative clauses per million noun phrases was higher in texts than conversations. Beyond overall frequency, different relative clause types tend to be used more or less often in written language compared with speech (Biber, 1988) and within written language, complex

grammar varies by genre – a novel compared with an academic article, for example (Biber, Conrad, & Reppen, 1998). These observations indicate that a range of discourse and contextual factors influence how adults use complex grammar when speaking and writing. While many studies have charted relative clause usage in children's early language development and related this to variations in spoken language input (e.g., Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010), how and when exposure to written language shapes grammatical development is not well understood.

Montag and MacDonald (2015) analysed a 2.4 million-word corpus of text written for school-age children, focusing on texts that children might read independently. They counted more relative clauses in this sample of book language than in child-directed speech. They also reported a higher ratio of passive to object relatives in written language than in speech, indicating greater complexity. Strikingly, books written for children contained a higher proportion of passive relatives than adult-to-adult conversation. Within children's books, the number of object relative clauses correlated positively with the intended age of each document indicating that as the texts increased in target age, so too did the number of relative clauses. What follows from these findings is the suggestion that learning to read and exposure to book language introduces substantial variation in the number and type of relative clauses children experience, well beyond the experience conferred by everyday conversation. Consistent with this suggestion, there is substantial variation in how well native-speaker adults comprehend complex grammar and these individual differences are associated with educational attainment (Dąbrowska, 2012; Dąbrowska & Street, 2006). Plausibly, these differences in spoken language comprehension might reflect, in part, differences in exposure to book language.

Montag and MacDonald (2015) analysed the content of books that children read independently. Importantly however, exposure to book language starts well before children can read for themselves. Shared reading – when a caregiver reads to a child – also provides opportunity to experience linguistic input that is quantitatively and qualitatively different to child-directed speech. Cameron-Faulkner and Noble (2013) analysed the language content of 20 picture books aimed at 2-year-olds and found that they contained more complex constructions than child-directed speech. This suggests that picture books provide enriched linguistic input, a conclusion supported by Montag's (2019) detailed analysis of complex grammar in a corpus of 100 picture books, also targeted at pre-schoolers. Montag found that sentences containing relative clauses (passives as well as subject, object, oblique and passive relative clauses) were much more common in picture books than child-directed speech. This suggests that the systematic grammatical differences in written vs. spoken language detailed in adult language are rooted in children's early language experience. This is an important observation given the huge variability in shared reading practices in the home. Logan, Justice, Yumus and Chaoarro-Moreno (2019) estimated that by the time children are 5 years old, those who have been read to five times a week will have experienced an additional 1.4 million words, compared to children not read to. While such observations have led to concerns about a substantial vocabulary gap associated with social disadvantage being firmly established by school entry, this variability in book language experience in the pre-school years also has serious implications for children's grammatical development.

Our first aim in this paper was to build on analyses of picture books (Cameron-Faulkner & Noble, 2013; Montag, 2019) and children's reading books (Montag & MacDonald, 2015) to quantify and directly compare the use of relative clauses across the two registers of book language and child-directed speech. In addition, our corpus of children's

reading books was sufficiently large to allow developmental slices to be made, based on the intended age of each book. We thus investigated whether relative clause usage changes with development, from books intended to be shared with pre-schoolers to those written for children to read independently from early through mid- and late childhood. Given the clear findings reported by Cameron-Faulkner & Noble (2013) and Montag (2019; Montag & MacDonald, 2015), we predicted that relative clauses would be more frequent in book language compared with child-directed speech; we also predicted that relative clause usage would increase as intended reading age increased. Alongside books for different ages, our corpus of children's reading books contained both fiction and nonfiction. This allowed us to compare relative clause usage across the two genres. In adult text, nonfiction is associated with more informational, technical and abstract language than fiction; compared to general fiction, academic prose and official documents contain more low frequency nouns, longer words and more prepositional phrases (Biber et al., 1998). Reading material used in early education tends to be narrative fiction. If nonfiction targeted at children contains richer and more complex language, there might be merit in developing nonfiction resources to support reading and language development (Kuhn, Rausch, Mccarty, Montgomery, & Rule, 2017; Lawrence, 2009).

Our second aim was to move beyond frequency counts of relative clause types to investigate lexical-syntactic patterns in children's book language. Different types of complex sentence are associated with certain types of words. People are highly sensitive to these lexical-syntactic combinations, as demonstrated by the sentence processing literature. For example, subject relative clauses are generally more frequent in English (Roland et al., 2007). In line with this, they tend to be easier to understand and produce than sentences that contain object relative clauses by adults (e.g., Gibson, 1998; Gordon, Hendrick, & Johnson, 2001,

2004; Grodner & Gibson, 2005; King & Just, 1991; Mak, Vonk, & Schriefers, 2006; Traxler, Morris, & Seely, 2002) and by children (e.g., Adani, 2011; Booth, MacWhinney, & Harasaki, 2000; Brandt, Kidd, Lieven, & Tomasello, 2009; Diessel & Tomasello, 2001; R. Macdonald, Brandt, Theakston, & Lieven, 2020). However, lexical-syntactic features such as the animacy of the noun phrase can alter the patterns seen in language corpora and this too is reflected in language processing patterns. There is a tendency for head nouns to be inanimate in object relative sentences (Roland et al., 2007). This correspondence between a lexical feature (animacy) and sentence structure (object relative) plays out in sentence processing, where the processing difficulty associated with object relative clauses is reduced when the head noun is inanimate (Betancort, Carreiras, & Sturt, 2009; Kidd, Brandt, Lieven, Tomasello, & Kidd, 2007; Macdonald, Brandt, Theakston, Lieven, & Serratrice, 2020; Mak, Vonk, & Schriefers, 2002; Traxler, Mason, Blozis, & Morris, 2005; Traxler et al., 2002). These and other types of lexical-syntactic variation and patterns in language experience have been related to processing differences in constraint-satisfaction accounts of sentence processing (Gennari & MacDonald, 2008, 2009; Hsiao & MacDonald, 2013; MacDonald, Pearlmutter, & Seidenberg, 1994). Much of this work has been situated in the adult sphere, informed by processing experiments with adults and testing sensitivity to usage statistics extracted from either adult language corpora or using estimates from child-directed speech. Given the differences between spoken and written language, however, there is a clear need to investigate developmental samples of book language. This will show how learning to read changes the nature of children's language experience and will pave the way to investigations of sentence processing that are more developmentally informed.

In this spirit, Montag (2019; Montag & MacDonald, 2015) tallied a range of lexical-syntactic combinations in children's books and found evidence of systematicity, beyond the

overall frequency counts of different type of relative clauses. This initial evidence bolsters the suggestion that exposure to book language provides critical linguistic input that shapes language development; in turn, this input should influence patterns of comprehension and production seen in older children and adults. Our aim was to replicate and build on Montag's work in several ways, using large developmental corpora. First, we compared lexical-syntactic combinations relevant to each relative clause type across child-directed speech and written language. Second, we compared two types of child-directed text – the language contained in picture books targeted at pre-schoolers and the language in books written for older children to read independently. Finally, and where relevant, we considered lexical-syntactic combinations across age and genre. Our aim throughout was to make links between the distributional patterns observed in children's book language and established findings in both the sentence processing and language acquisition literatures.

Method

Description of corpora

We analysed three different corpora. Two of these comprised child-directed text and of these, one contained books written primarily for pre-school children to hear in the context of shared reading with caregivers and the other books for independent reading by older children. The third corpus contained child-directed speech targeted at pre-school children.

(i) ***Picture book corpus*****.** This newly constructed corpus (see also Dawson et al., 2021) comprises 160 children's fiction books with a total word count of 316,711. These books were selected to be representative of the type of reading material children encounter in shared reading contexts in the UK. To this end, we generated an initial list of titles with a target age

range of 0-7 years from a combination of retailer bestseller lists and recommendations from literacy charities, book review sites, and teachers. The final list of purchased books (Appendix A) included the titles that were cited most frequently across these sources. Most books in the corpus were picture books, but a small number of longer texts that might be read to young children were also included (e.g., *The BFG*). The content of each book was transcribed as plain text files. We included text that appeared in illustrations and appendages (for example, text in speech bubbles) in the transcription on the basis that caregivers would likely read these words aloud in addition to the main body of text.

(ii) ***Reading book corpus*****.** Analyses were based on the reading component of the Oxford Children's Corpus, developed and held by Oxford University Press. This dynamic and growing corpus contains language written for 5-14 year-old children. We sampled the corpus at a size of 13,154 documents (about 34 million words) spanning fiction, nonfiction, curriculum materials and children's websites. For some texts, Key Stage metadata provided an indication of developmental level. Key Stage refers to age bands in the education system of England and Wales (Key Stage 1: 5-7 years; Key Stage 2: 7-11 years; Key Stage 3: 11-14 years).

(iii) Child-directed speech. Our corpus of child-directed speech was generated from 10 corpora in the English-UK section of the CHILDES database (MacWhinney, 2000). The sample comprised all suitable corpora from this collection, with the exception of those that focused on specific populations (e.g., children with language impairments). The final set of 10 corpora (see Appendix B) contained transcripts of interactions between 190 different children aged from 6 weeks to 6 years and their caregivers, siblings, other family members and researchers. Recordings took place across a variety of contexts, but typically involved

structured and free play activities, as well as everyday routines such as mealtimes and bedtimes. Across all recordings, utterances produced by the child were filtered out, such that the final dataset comprised only talk directed to the child, totalling 3,771,352 words.

Identification and classification of relative clauses

To extract and analyse relative clauses from each corpus, we first parsed the content from each of the three corpora into a syntactically searchable format using the Berkeley Neural Parser (Kitaev & Klein, 2018) implemented in Python. The parser is attested with high classification accuracy, with the F1 score being 95.13 with pre-training (Kitaev & Klein, 2018).This generated constituency parser trees that represented the hierarchical syntactic structure of each sentence. We used the software Tregex (Levy & Andrew, 2006) to extract the major types of relative clauses. Tregex utilizes regular expressions to match patterns in the trees; the expressions used are provided in Appendix C.

We focused on the types of relative clauses that previous psycholinguistic literature consider canonical, which are those that modify an overt noun phrase as the antecedent. There are four main types of relative clauses: subject relative clauses, object relative clauses, oblique relative clauses, and passive relative clauses (see Table 1 for examples of each type of relative clauses). Subject relative clauses modify an entity that performs an action with or without an affectee. That is to say, a subject relative clause contains either a transitive or an intransitive verb. Object relative clauses were classed as modifying a direct object or an indirect object that an agent performed an action on. The relative pronoun, like "*which*", "*who(m)*", "*that*", can be omitted. Oblique relative clauses modify nouns that are neither subjects nor objects; this type of relative clause usually ends with a preposition. Note that the automatic parser did not always differentiate oblique from object relative clauses that contain

phrasal verbs (i.e., using the same pattern would sometimes extract both types of relative clauses). For example, a pattern that extracted an oblique relative clause like "*the crayon (that) he drew with*", would also extract object relative clauses that contained a phrasal verb, as in "*the income (that) she relied on*". Given that object relatives and oblique relatives have been treated as separate categories in previous studies (e.g., Montag, 2019), and that the cases of object relative clauses with phrasal verbs were few, we labelled all cases that ended with a preposition as oblique relative clauses. The fourth category of relative clause, passive relatives, are structurally similar to subject relative clauses but the position of the agent and patient is reversed, such that the patient is the head noun of the relative clause and the agent is specified through the "by-phrase". Agent information can be omitted entirely by dropping the by-phrase. Some linguists argue that certain verbs in past participle form should instead be considered adjective (e.g. dressed, named), such that the phrases "*the girl dressed in white", "the girl named Jane*" were not passive relative clauses. However, given that Roland et al. (2007; see their Table 5) included them as members of the passive relative clause category, and that the automatic parser was not discriminatory of such cases, we included them under passive relatives.

Table 1

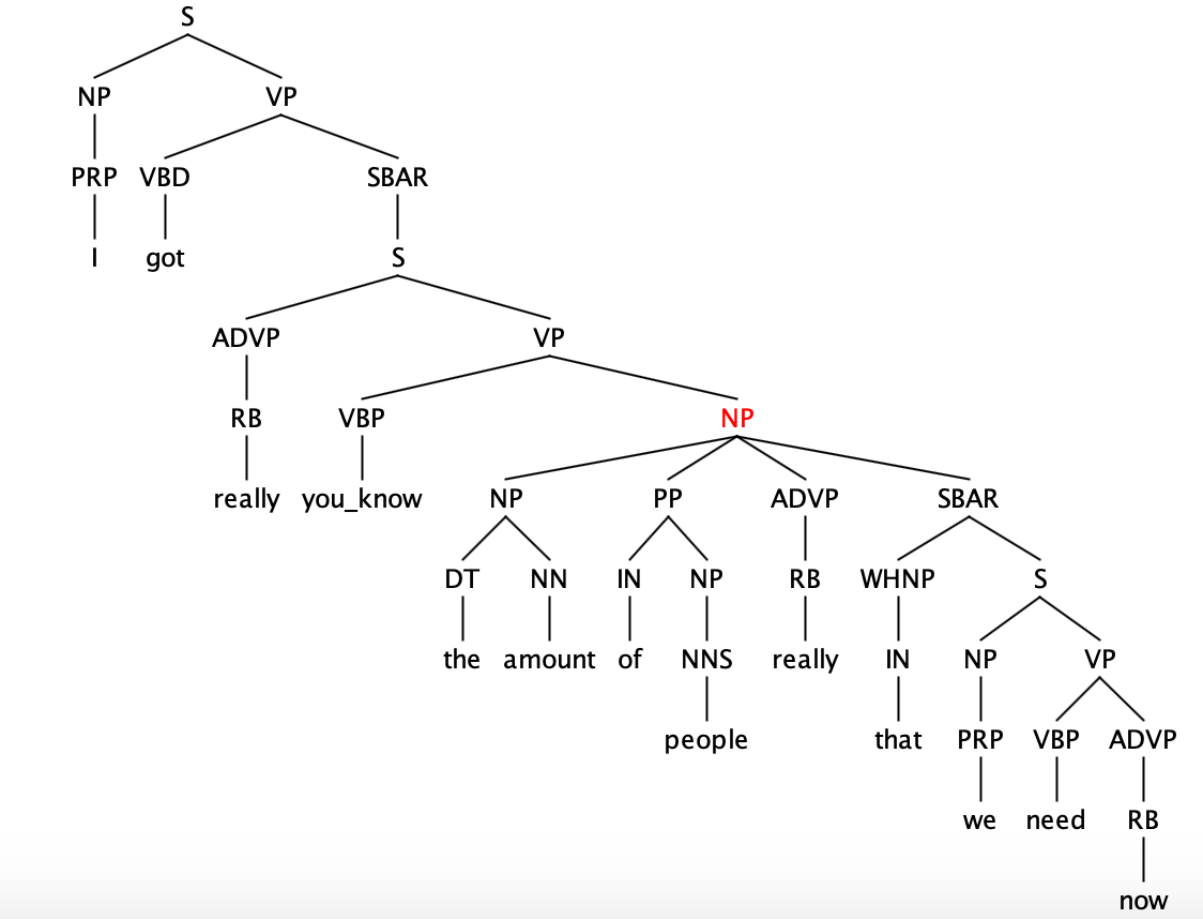
Examples of the four categories of relatives clauses examined in this study. Parentheses indicate elements that can be omitted.

Relative clause type	**Examples**
Subject relative clause	*"the bridge which spanned the chasm"* (transitive) *"the boy who jumped*" (intransitive)
Object relative clause	*"the goals (that) the world leaders set*" (direct object) *"the boy (that) she gave the book to"* (indirect object)
Oblique relative clauses	*"the crayons (that) you draw with"* object relative clauses with phrasal verbs were also included, like *"the income (that) she relied on"*
Passive relative clause	*"the part (that is) lit up (by the sun)"*

Automated analysis risks issues with accuracy. Sentences that contain highly complex structures, ungrammatical fragments, or interrupted phrases can be difficult for the parser to detect. This means that we may have missed instances that should have been included as target structures, or mis-captured instances that are not true examples of the target structures. This might be especially an issue for speech data, as observed in adult corpora (Roland et al., 2007). Figure 1 shows an incorrectly parsed example, extracted from CHILDES, which contains colloquial intervening phrases, such as "*you know*", and "*really*". In this instance, the parser incorrectly parsed the noun phrase marked in red as the direct object of "*you know*". In addition, although we devised Tregex patterns to capture relative clauses, it is possible that they also captured irrelevant sentence structures, or simply missed the target structures.

Figure 1.

An example parse tree of a mis-parsed sentence containing interjections like "you know" extracted from CHILDES



Given these issues with automatic coding, it is important to establish its accuracy and understand the nature of the errors it returns. To this end, we randomly sampled 1000 sentences from each of the three corpora. A research assistant with advanced training in linguistics hand-coded whether any of these sentences contained the four types of relative clauses. The coder and the first author reached 100% agreement on the criteria for coding each relative clause type, informed by linguistic theories (meaning that both people agreed on the coding judgments across the 3000 sentences). We then compared the hand-coding with the results generated by the automated procedure. Table 2 lists the raw number of relative clauses identified in each corpus based on the two methods. It also includes values for precision, recall and F measure, as commonly used in the Natural Language Processing literature to access model performance (Jurafsky & Martin, 2000). Precision refers to the

percentage of correctly identified cases out of all identified cases (percentage of true positive out of all true positives and false positives). Recall represents the percentage of all identified cases that were correctly identified (percentage of true positives out of all true positives and false negatives). The F measure is the weighted harmonic mean of precision and recall, operationalised as 2 * precision * recall/ (precision + recall).

Table 2.

Raw frequency of relative clauses, precision, recall and F values across corpus using automated extraction compared to manual coding on the 1000 random sentences sampled from each corpus. Inside the parentheses contains the number of false positives and false negatives of machine identification.

machine identified/manual coding (false positives, false negatives)	**Child-directed speech**	**Picture Books**	**Reading books**	**Total**	**Precision**	**Recall**	**F**
Subject relative	8/8 (0, 0)	25/25 (0, 0)	45/45 (0, 0)	78/79 (0, 0)	100%	100%	100%
Object relative	3/3 (1, 1)	19/18 (3, 2)	24/23 (2, 1)	46/43 (6, 4)	87%	91%	89%
Oblique relative	0/0 (0, 0)	2/3 (0, 1)	7/8 (0, 1)	9/11 (0, 2)	100%	82%	90%
Passive relative	1/1 (0, 0)	7/6 (1, 0)	20/20 (0, 0)	28/26 (1, 0)	96%	100%	98%
All relative clauses	12/12 (1, 1)	53/52 (4, 3)	96/96 (2, 2)	161/159 (7, 6)	96%	93%	94%

It is clear from Table 2 that the overall frequency of relative clauses was low, comprising about 5% of all data. Child-directed speech had the fewest number of relative clauses overall, followed by picture books, and in turn by reading books, where relative clauses were most frequent. This is consistent with our expectation that complex syntax is more common in written language and that this increases as intended age builds. The automated procedure had high accuracy, with high precision, recall and F measure. For items

that generated disagreement between the automated and manual methods, we highlight some findings here; a more detailed and systematic error analysis is provided in Appendix D.

There were fewer disagreements between hand- and automated coding for child-direct speech than for written language. This may be the direct result of there being fewer relative clauses in speech (11 occurrences, compared to 52 in picture books and 96 in reading books); speech also comprised shorter sentences (average 3.9 words in each sentence, compared to 10.9 words in picture books and 13.5 in reading books) and less complex structure (only 16% of sentences contained more than two lexical verbs, compared to 55% in picture books and 57% in reading books) (see supplementary materialsl). By relative clause type, we note the relatively higher rate of disagreement for object relative clauses in the reduced form (when relative pronoun was omitted) compared to other types. Several disagreements (5 out of all 9 disagreements for object relatives) belonged to cases like "*the way you talk*" and "*the moment he arrived*", where the modified noun was not the object of the relative clause. These are termed relative adverbial clauses, and oftentimes the head noun can be replaced by a relative adverb (e.g. "*I like <u>the way</u> you talk*" can be rephrased as "*I like <u>how</u> you talk*"). Given that this type of clause are not distinguishable structurally from regular object relative clauses, we included them as members of the object relative category. Other cases of disagreement for object relative clauses originated from the automatic parser not segmenting the clause boundary correctly, especially when the relative clause pronoun was omitted. For example, in a reduced object relative clause like "*the book grandmere used*" (the reduced form of "*the book that grandmere used*"), the parser treated "*the book grandmere*" as a single noun phrase instead of two, perhaps because of the unusual spelling of the loan word "grandmere". Most other disagreements could also be attributed to parsing errors by the automatic parser. These include mistakes in parsing long run-on sentences, head nouns that were ambiguous in grammatical function (e.g. "*present*" can be a noun, verb and an

adjective), or simply mistakes that did not have a clear explanation (e.g. the verb "wear" in "*the blue flowers I always wear*" was parsed as a punctuation). Similar patterns of errors were observed for oblique relative clauses (e.g. in "*there wasn't much Hubert didn't excel at*", "much Hubert" was parsed as a single NP). In summary, the written corpora and object/oblique relative clause types were more susceptible to parsing errors in automatic coding. The error analysis indicates that this is likely due to a number of factors, including longer and more complex sentences and the presence of unusual words (e.g. loan words, coined words) written text contained (see Appendix D for detailed description and analysis).

Having established the estimated accuracy of our automated procedures for extracting and classifying relative clauses from each corpus alongside manual coding, we next derived the overall frequency of different types and subtypes of relative clauses in the three corpora in Analysis 1 before considering the lexical-syntactic distributions that characterise each type of relative clauses in Analysis 2.

Analysis 1: The frequency of relative clause types in children's book language

Our first aim was to replicate and extend at scale earlier work comparing complex grammar in children's books with child-directed speech (e.g. Caremon-Faulker & Noble, 2003; Montag, 2019; Montag & MacDonald, 2015). We began by computing the number of relative clauses of each type across the three corpora. Based on previous findings, we expected all four categories of relative clauses to be more common in book language than spoken language. We then used the metadata available in the Oxford Children's Corpus to slice its content by targeted Key Stage, and by genre. This allowed us to consider the distribution of relative clause types as intended age increased, and in fiction vs. nonfiction text.

Results and Discussion

(i) Frequency and distribution of relative clause types

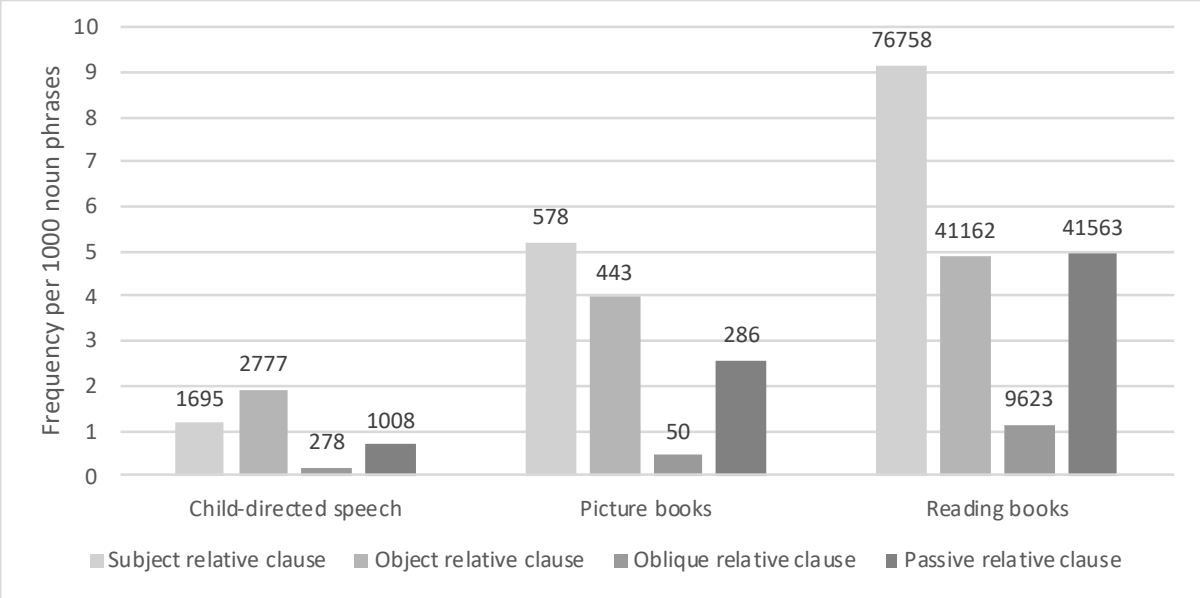
Due to the unequal size of the corpora and following Roland et al. (2007), we first normalized the frequency of each type of relative clause by the number of total noun phrases in that corpus (a relative clause can only modify noun phrases). There were 1,451,545 noun phrases in total in the child-directed corpus, 110,863 in the picture book corpus, and 8,380,889 in the reading book corpus. The proportion of relative clauses of each type out of all noun phrases was then multiplied by 1000 to make the value more interpretable. The distribution of relative clauses per 1000 noun phrases across corpora is depicted in Figure 2, along with the raw count. Subject relatives were most frequent across corpora (15.54 occurrences per 1000 noun phrases, 79031 raw occurrences in total). Object relative clauses were slightly less frequent (10.82 occurrences per 1000 noun phrases, 44382 raw occurrences in total), followed by passive relative clauses (8.23 occurrences per 1000 noun phrases, 42857 raw occurrences in total). Oblique relative clauses were least frequent (1.79 occurrences per 1000 noun phrases, 9951 raw occurrences in total).

As is clear from Figure 2, all types of relative clauses were less frequent in child-directed speech than in either sample of book language. The contrast between picture books and child-directed speech is particularly informative as both contain language targeted primarily at pre-schoolers. Even when the age of the child is comparable, there were more relative clauses in book language than spoken language (12.24 vs. 3.97 relative clauses per 1000 noun phrases). Across the two types of book language, picture books contained fewer relative clauses than books written for children to read themselves (12.24 vs. 20.18 relative clauses per 1000 noun phrases). The pattern of relative frequency across the four different

types, however, was similar between the two book language corpora, with subject relatives most common. In child-directed speech, object relatives were most frequent. In all three corpora, oblique relative clauses were the rarest among all types.

Figure 2.

The frequency distribution of the three types of relative clause per 1000 noun phrases in child-directed speech, children's picture books, and children's reading books. *Raw frequency is shown as labels.*

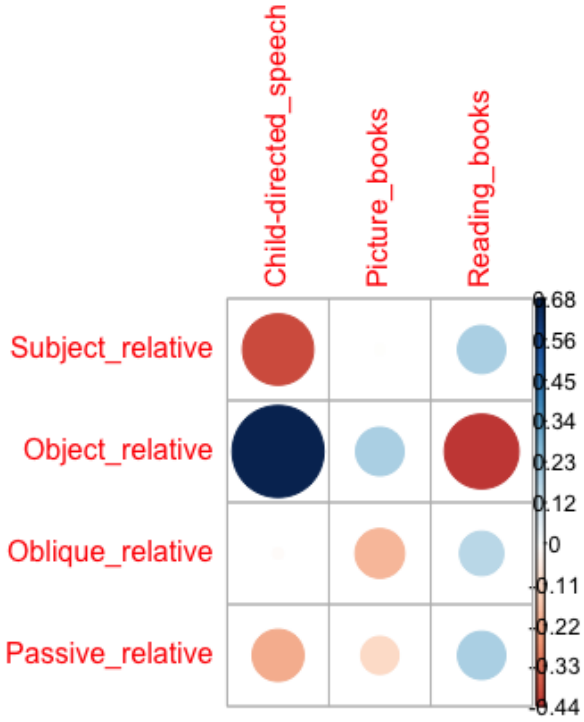


We next asked whether specific corpora featured certain types of relative clauses more than the others. Using a Pearson's Chi-squared test, we tested whether the two variables – corpus and relative clause type – were associated. Figure 3 visualises the Pearson residuals that measure the relative association between the four types of relative clause across the three corpora. A positive association indicates that the type of relative clauses was characteristic of the text, indicated in blue. A negative association, shown in red, indicates that the relative clause type was not representative of the language contained in that corpus. The darker and larger the circle, the stronger the (dis)association. Independence between corpus and relative

clause types could not be established (χ^2 (6) = 1.04, p < .98), suggesting that types of relative clauses were associated with specific corpora. Figure 3 shows that object relative clauses were strongly associated with child-directed speech, and negatively associated with texts for independent reading. Subject relatives, on the other hand, were negatively associated with child-directed speech. Examining closely, we found that object relatives occurred in child-directed speech mostly resembled such cases as "*all I have*" and "*nothing I can do*". The relative clause pronoun was omitted, the modified noun was unspecified or indefinite, such as "*all*", "*anything*" or "*nothing*", and the agent was a pronoun, dominantly being "*I*" or "*you*" given the interactive nature of speech and the focus of the caretaker on the child. For picture books, although such reduced object relatives were still frequent, the modified nouns became more specified, taking the form of full noun phrases (e.g. "*the cake she had made*"). For both picture books and reading books, subject relatives were overall most frequent, particularly those with intransitive verbs like "*the boy who ran away*". We discuss these lexical structural co-occurrences in more detail in Analysis 2.

Figure 3.

Correlation plot of Chi-square residuals of relative clause frequency by type and corpus. Blue indicates a positive correlation and red a negative correlation. Larger circles indicate stronger (dis)association.



(ii) Analyses by intended reading age

Having established that relative clauses are more frequent in book language than child-directed speech, we used the metadata available in the Oxford Children's Corpus to examine developmental trends in the distribution of relative clauses, as its content becomes more targeted towards older children. Where metadata was available, most material fell within Key Stages 1 to 3. Splitting into sub-corpora, Key Stage 1 (5-7 years) contained 2.2 million words (567,409 noun phrases), Key Stage 2 (7-11 years) contained 18.5 million words (4,968,757 noun phrases), and Key Stage 3 (11-14 years) contained 9.7 million words (2,844,723 noun phrases). Documents without Key Stage metadata were excluded from the following analyses.

Figure 4 shows the frequency of the four types of relative clauses across each of the three developmental windows. Once again, frequency is plotted as the number of relative clauses per 1000 noun phrases normalised for the size of each sub-corpus and raw frequency

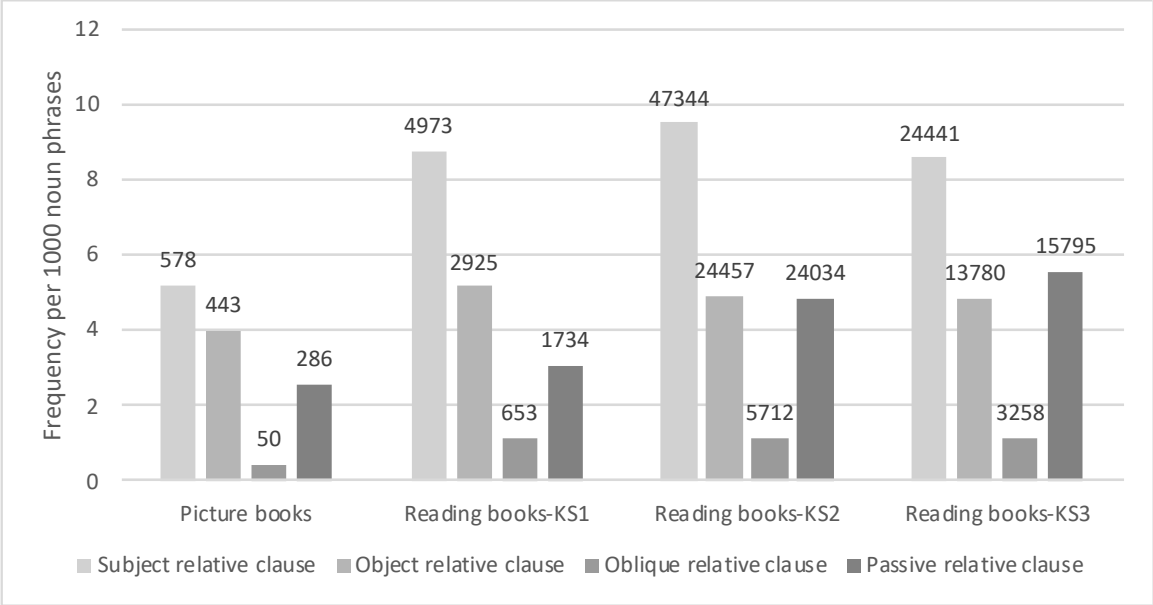
is provided as a label; for comparison, data are plotted alongside data from the picture book corpus (aimed at pre-schoolers). As to be expected given the overall analyses reported above, all four types of relative clause were more frequent in the reading books corpus than the picture book corpus. Even those texts targeted at Key Stage 1 children contained more relative clauses than picture books (18.13 vs. 12.24 relative clauses per 1000 noun phrases).

The frequency of all relative clauses was similar across the Key Stage bands in the reading book corpus, except for passive relative clauses. These showed a more stepwise increase in frequency with developmental level (all pairwise comparisons of proportions of passive relative clause frequency between developmental levels were significant, $p < .01$). Passives are difficult for children because of the unusual word order and thematic role alignment (Boyle, Lindell, & Kidd, 2013; Montgomery & Evans, 2009). Plausibly, reading experience allows children to master this structure gradually over time. In the discourse sense, the passive voice is also more impersonal and neutral (Ding, 2002; Rundblad, 2007; Tarone, Dwyer, Gillette, & Icke, 1998), especially when the identity of the author or the doer of action is masked through agentless passives. This discourse function operates particularly in nonfiction, which constituted higher proportion of text as Key Stage increased (none in the picture book corpus, 1% in Key Stage 1, 18% in Key Stage 2 and 25% in Key Stage 3). For passive relatives used in picture books, which constituted entirely fiction, many instances indicated names, e.g. "*a soft brown toy called Dogger*" (26% of all reduced passives without the by-phrase, compared to 15% in the reading books overall). We discuss this in more detail in the next section on genre and later in Analysis 2 on lexical syntactic co-occurrences, but in the meantime, it might explain why passive relative clauses are rare in text written for younger children. In all the book corpora, subject relative clauses were the most frequent type. This indicates that this type of relative clause is most characteristic of written language.

However, although there was a significant increase of subject relatives from picture books to Key Stage 1 reading books (χ^2 (1) = 6.92, p = .009), no upward trend was observed with Key Stage later on. Object relatives were common in picture books and reading books for Key Stage 1, but were less so for text targeted at older children. Oblique relatives remained rare across developmental stages.

Figure 4.

The frequency distribution of the three types of relative clause by corpus as a function of intended developmental level. Raw frequency is indicated as labels.



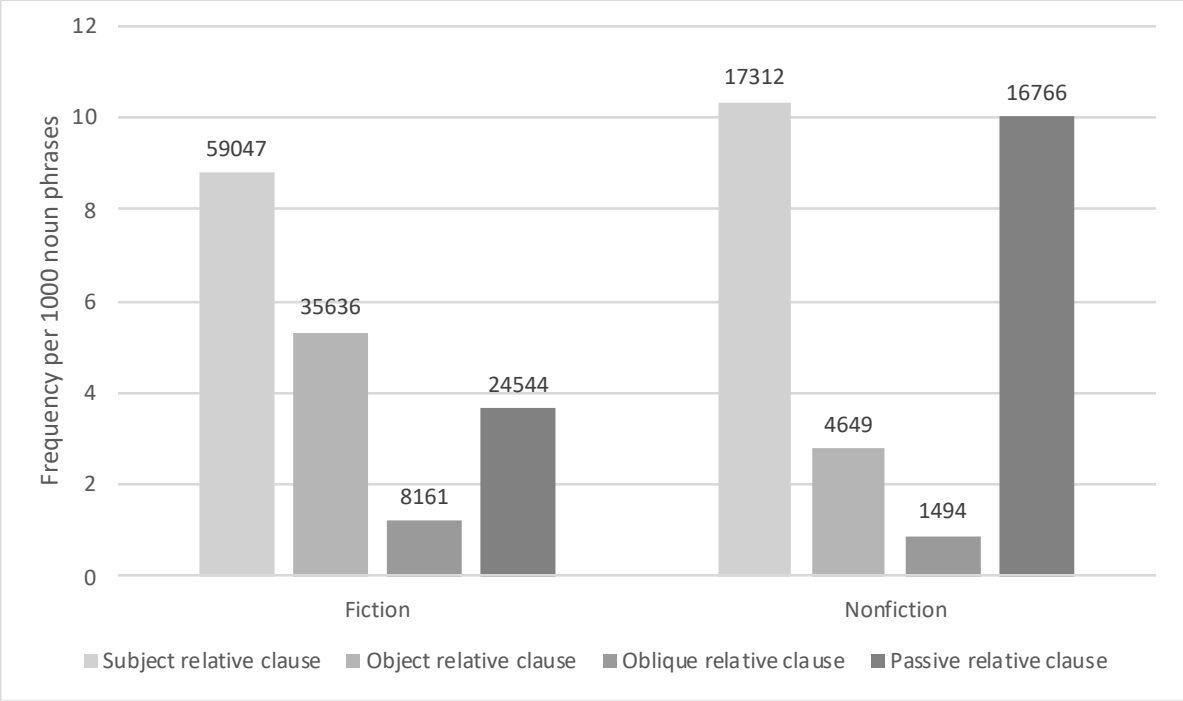
(iii) Analyses by genre

The reading book corpus contains a large number and variety of texts. The metadata permit a division between fiction and nonfiction, with around 80% of the corpus being fiction. This allowed us to examine relative clause usage across genre. Note that we did not include the picture book corpus in these analyses as it contains no nonfiction texts. Figure 5 shows the raw frequency and the frequency per 1000 noun phrases for each relative clause type, with frequency normalised for the size of each sub-corpus. Overall, nonfiction contained significantly more relative clauses than fiction (24 compared to 19 per 1000 noun

phrases, 127388 vs. 40221 in raw frequency, χ^2 (1) = 1723, $p < .001$). Testing within each relative clause type, nonfiction contained many more passive relative clauses (10.01 vs 3.66 per 1000 noun phrases, 16766 vs. 24544 in raw frequency, χ^2 (1) = 11019, $p < .001$) than fiction. Subject relative clauses were also more common in nonfiction than fiction (10.34 vs 8.81 per 1000 noun phrases, 17312 vs. 59047 in raw frequency, χ^2 (1) = 348, $p < .001$), whereas object relatives were more common in fiction (5.32 vs 2.78 per 1000 noun phrases, 35636 vs. 4649 in raw frequency, χ^2 (1) = 1804, $p < .001$), as well as oblique relative clauses 1.22 vs 0.89 per 1000 noun phrases, 8161 vs. 1494 in raw frequency, χ^2 (1) = 123, $p < .001$). These findings indicate that passive relatives and subject relatives might be more instrumental for expository nonfiction text, whose purpose is to provide informational content about a topic (e.g. "*A migrant is a person who moves to another country*", "*The Ancient Romans played a game a bit like golf , using sticks and a leather ball stuffed with feathers*"). The dominance of passive relative clauses in nonfiction again reflects the discourse requirement of such genre being more impersonal and neutral. In contrast, fictional narratives that describe relationships between characters and objects employ more object and oblique relative clauses (e.g. "*He could make Menie hear everything he said*", "*The particular straw I 'm clutching at is gold*".)

Figure 5.

Frequency per 1000 noun phrases, as well as raw frequency, of each type of relative clauses by genre across children's reading books



To summarise the findings of Analysis 1, book language contains more relative clauses than child-directed speech. This remains the case when the comparison is restricted to picture books typically read to pre-school children. In turn, books written for children to read independently contain more relative clauses than picture books. This pattern was evident in the analysis of Key Stage 1 books written for 5-7 year-old children, indicating that it is a characteristic of book language from the early stages of reading development. We also saw differences in the type and distribution of relative clause across corpora and sub-corpora. Object relatives were the most common type in child-directed speech but were less common in book language, and in nonfiction in particular. In contrast, passive relatives were rare in child-directed speech but became gradually more common in texts for older children, and in nonfiction. Subject relative clauses occurred more often in picture books for pre-schoolers than speech directed to children of similar age; they were more frequent still in books for independent reading, and in nonfiction. Although oblique relative clauses were the rarest type across all corpora, they were more common in books than in speech, and in fiction than

nonfiction. Taken together, these frequency counts and cross-corpus comparisons show that book language provides children with exposure to variations in complex grammar from the outset and as targeted developmental level of text increases, so too does the amount and nature of complex grammar.

Analysis 2: Lexical-syntactic variation within relative clause types

Different types of complex sentence are associated with certain types of words in ways that are systematic and predictable (Roland et al., 2007). To fully capture relative clause usage in book language, it is therefore important to move beyond frequency counts and consider the nature of lexical-syntactic distributions. In this section, we took a detailed look at each type of relative clause. For each, we started with a lexical-syntactic feature that characterises adult language input and sentence processing (e.g. noun animacy, verb transitivity, pronoun vs. full noun status) and asked how it is represented in book language. Where appropriate, we compared book language with child-directed speech, and investigated developmental trends and differences across genre. Note that in this analysis, we classed object relative clauses and oblique relative clauses together, reflecting that the two types are not are not clearly defined in the processing literature and have similar processing profiles. Furthermore, this represents a conservative approach as our automatic parser was not always able to differentiate the two, as discussed earlier.

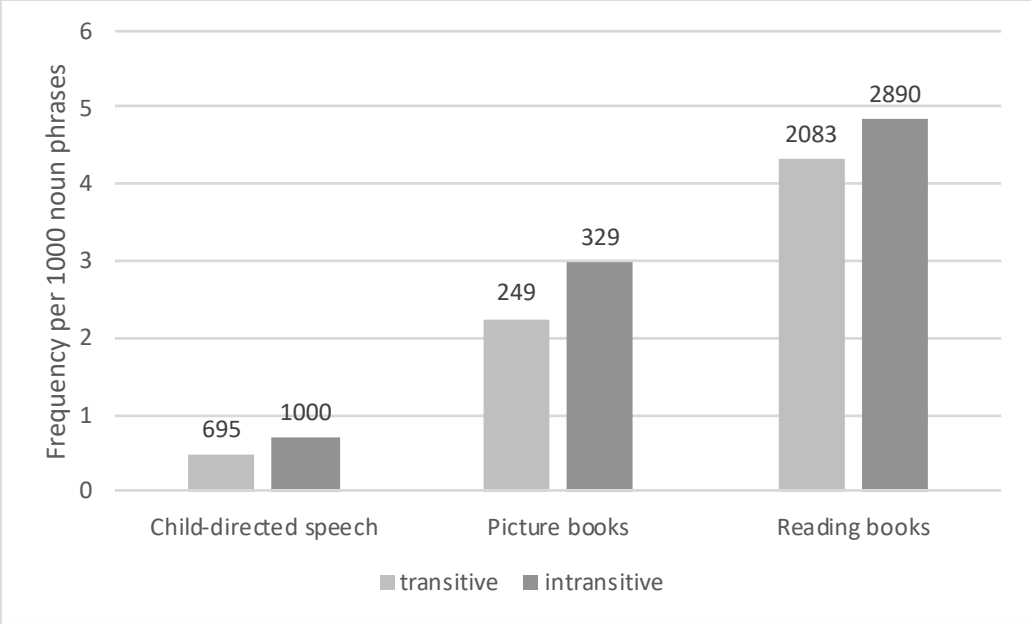
(i) Subject relative clauses

There are two main types of subject relative clause: one that takes an object noun phrase and one that does not. Transitive verbs take direct objects, therefore creating a more

complex argument structure, compared to intransitive verbs which cannot. Figure 6 shows the frequency in raw and normalised terms of subject relatives as a function of transitivity in each corpus. Both types were more common in book language, consistent with the function of a subject relative clause (expansion of the information on a noun in focus) being more required in text relative to speech, while the proportion of intransitive relative clauses was higher overall than the transitive type. This replicates previous findings that saw lower frequency of transitive subject relatives compared to intransitive ones in picture book corpus (Montag, 2019) and in children's spontaneous speech (Diessel, 2004), as well as lower performance on transitive subject relatives compared to intransitive ones in experimental data (Diessel & Tomasello, 2005).

Figure 6.

Raw frequency and normalised frequency of transitive and intransitive subject relative clauses in child-directed speech and book language.

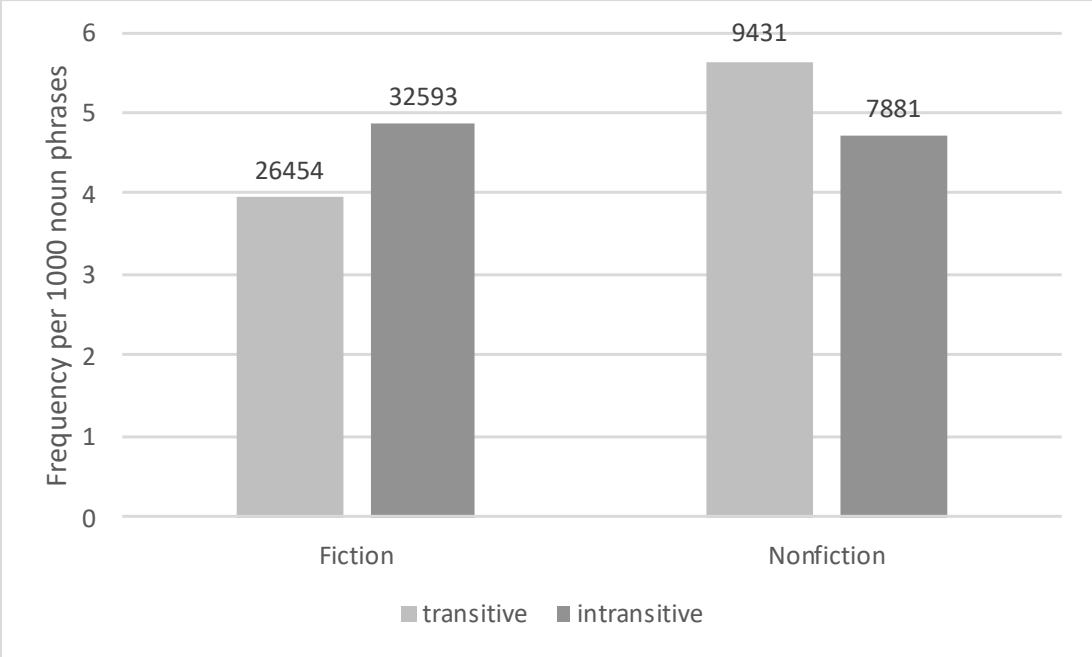


We explored this changing distribution in two ways. First, we charted the proportion of light verbs, that is, verbs with low informational value (e.g., *be, have, get, go, run*; e.g. "*the boy who is five*"). Analyses of adult speech show that light verbs are commonly produced in subject relative clauses, and that this tendency is higher in spoken language than written language. Roland et al. (2007) reported that 29% of subject relative clauses in spoken language (extracted from the Switchboard corpus) contained '*be*' verbs, and this extended to 50% when other light verbs were included. The comparable figure in written language (extracted from the Brown corpus) was only 17%. Roland et al. further observed that in spoken language, many instances of light verbs were associated with the difficulty of producing certain lexical items (e.g. "*the people who run the prison*" to replace "*the wardens*"). Similar findings have been reported for children's own speech. Diessel (2004) found that about a third of subject relatives in a speech corpus produced by two English-speaking children contained the copula "be" (e.g. "*some apples that were sweet*"). In adult written language, however, Roland et al. found that subject relative clauses were more likely to provide additional information. We therefore predicted that the proportion of subject relative clauses containing "*be*" verbs would be lower in book language than child-directed speech. The hypothesis was confirmed: 34% of all subject relative clauses in child-directed speech contained "*be*" verbs, compared to 19% in picture books and 15% in books for independent reading. Within book language, we hypothesised that the discourse requirements, especially for nonfiction, would lead it to contain more transitive subject relative clauses: as these require a more complex argument structure, they should be less likely to involve light verbs such as "*be*". As shown in Figure 7, nonfiction texts contained more subject relative clauses overall in normalised frequency, and more transitive subject relative clauses than fiction (χ^2 (1) = 503, $p < .001$), consistent with our hypothesis. However, within the same data, the number of "*be*" verbs was at 15% for fiction and 16% for

nonfiction. This difference was significant (χ^2 (1) = 9.16, p < .002). It might be that other light verbs (e.g., *have, get, go*) were more common in fiction, thus masking the difference across genre we expected to see.

Figure 7.

Frequency per 1000 noun phrases of subject relative clauses with transitive and intransitive verbs in the reading book corpus by genre



Finally for subject relative clauses, we examined the lexical factor of animacy. We predicted that transitive subject relatives would most often modify animate nouns and contain an inanimate object as the embedded noun, as in "***the boy*** *that read* ***a book***". To test this hypothesis, we selected at random 100 transitive subject relative clauses from each of the three corpora and hand coded the animacy of nouns. We classed nouns as animate if the entity it refers to possesses agency and volition in order to perform an action; entities without these properties were considered inanimate (Hundt, 2004). Consistent with our prediction,

58% of head nouns were animate (61% for child-directed speech, 69% for picture books, 44% for reading books) and 72% of the embedded object nouns were inanimate (83% for child-directed speech, 64% for picture books, 69% for reading text). However, among intransitive subject relatives, only around 38% of the modified nouns were animate (34% for child-directed speech, 41% for picture books, 39% for reading text). Closer examination of the intransitive instances revealed that those with animate heads tended to describe an action (e.g. "*the boy who turned around*"), whereas those with inanimate heads usually involved generic or abstract verbs (e.g. "*the footprints that ran across the floor*").

(ii) Object relative clauses and oblique relative clauses

According to classic studies in the adult sentence processing literature, object relative clauses are difficult to process as they place high demands on working memory (Chomsky & Miller, 1963). Yet in our analyses of child-directed speech (see Figure 2) and elsewhere (Montag, 2019), object relative clauses are frequent. This leads to an apparent paradox: there is evidence of processing difficulty despite object relatives being high in frequency. This paradox may be more apparent than real, however, given that the type of object relative clauses often used in the sentence processing literature tends to be rare in the linguistic environment (e.g., Gennari & MacDonald, 2008). Our three corpora offered an opportunity to investigate the nature of object relative clauses through development, and across spoken and written language experience.

We also considered oblique relative clauses. These have been studied less in the processing literature than object relative clauses (Kim, 2016). In fact, the definition of oblique relative clauses is not clear or consistent. Some definitions are more lax and count all

relative clauses that end with a preposition (e.g. "*income that the she relied on", "the boy that the girl played with"*) as obliques. Other definitions are more constrained and include in this category only those relative clauses with a preposition not tightly connected to the verb, hence the term oblique (e.g., the verb "rely" cannot be used alone without the preposition "on", and therefore "*the boy that the girl relied on*" is an object relative clause but "*the boy that the girl played with*" is an oblique relative clause). On top of the linguistic complexity and definitional issues, the constituency parser we used (like other dependency parsers currently available) was not able to distinguish oblique relative clauses from object relative clauses that ended with a preposition. We therefore took the lax approach in this analysis and categorised any relative clause ending with a preposition as an oblique relative clause.

Our analyses were informed by the sentence processing literature. The type of canonical object relative used in psycholinguistic studies tends to involve a relative clause pronoun (e.g., *that, which, who* or *whom*) as well as full noun phrases that also denote animate entities (e.g., "*the official that the reporter criticised*"). These are difficult to process. They become less difficult when the relative clause pronoun is absent, the head noun is inanimate and the embedded noun is a pronoun, as in "*the book I read*" (Gordon et al., 2001, 2004; Traxler et al., 2002). Similar distributional patterns appear to characterise oblique relative clauses. Gennari and MacDonald (2008) found that when provided with a sentence fragment "The N that the" or "the N that the N", participants tended to continue the sentence to form an oblique relative clause, especially when the head noun was inanimate, as in "*The play the actor performed in*". With these observations as a backdrop, we examined object and oblique relative clauses together in children's language, looking closely at those that omit the relative pronoun, contain inanimate head nouns and include a referring pronoun.

As shown earlier in Figure 2, the frequency of object relative clauses (e.g. "*the book I read*") was more common overall across the three corpora than the oblique type (e.g. "*the crayon he drew with"*). It is also clear that both types were more common in book language than child-directed speech; within book language, both types of object relative were more common in reading books than picture books. The proportion of these sentences that included relative clause pronoun omission was high, using the markers "*that*", "*who*", "*which*" and "*whom*". Omission was common, and it was more likely to be seen in object relative clauses (73%) than oblique relative clauses (52%). Omission was also less likely in reading books for both types of relative clauses than both picture books and child-directed speech: 64% of the object relative clauses did not contain the optional relative pronoun in the reading books, compared to 77% in both picture books and child-directed speech. The rate of omission was even lower for oblique relative clauses at 31% in the reading books, compared to 68% in picture books and 56% in child-directed speech. This might reflect a developmental transition in written language as it becomes more formal in style.

Turning to animacy, we next asked whether object relative clauses and oblique relative clauses were more likely to modify inanimate entities than animate entities. We randomly sampled 100 sentences from each of the three corpora that contained these relative clauses and manually coded for the animacy of the head noun. Note that there were only 42 oblique relative clauses in the picture book corpus. The majority of the head nouns being modified by object relatives and oblique relatives were inanimate, at 96% and 91%, respectively. This was also true when the data was split by corpus: for object relatives, 98% in child-directed speech, 97% in picture books and 94% in reading books had inanimate heads; for oblique relatives, 93% in child-directed speech, 88% in picture books, and 90% in reading books had inanimate heads. This confirms that the lexical-syntactic pairing of object

relative clauses with inanimate nouns is of high frequency in language input from an early age.

Our final analysis of object and oblique relative clauses considered the prevalence of a pronoun being in the embedded subject position across the three corpora. As noted above, sentence comprehension tends to be easier when there is a pronoun (e.g., "*the book **he** read*") rather than a full noun phrase (e.g., "*the book **the little boy** read*") (Gordon et al., 2004). We therefore anticipated pronouns would be prevalent. In line with this, the percentage of pronoun use in the subject position was high based on the same 100 object relative clauses sentences selected at random from each corpus, as described above, with 74% across all corpora (91% in child-directed speech, 65% in picture books and 65% in reading books). Similar trends were observed with oblique relatives, with a total of 77% of pronoun use across corpus (84% in child-directed speech, 69% in picture books and 74% in reading books). We further coded the animacy of these pronouns based on the hypothesis that they would mostly denote animate entities which acted upon inanimate head nouns. For pronouns whose animacy was not readily obvious (e.g., *it, they*), we used the context to determine animacy. Across all the three corpora, when the subject was a pronoun it almost always referred to an animate entity, at 99% for both object and oblique relatives. This was also the case when the subject was a full noun phrase, at 82% for object relatives and 80% for oblique relatives. Based on noun animacy and pronoun status of the relative clause subject, we can state that the distributional properties of object and oblique relative clauses tend to have an inanimate head noun, an embedded relative clause subject that is animate and a pronoun, as in *"the book I read"* and *"the crayon he drew with*". We crossed these three factors and checked that this type was the most common in the 100 random samples of object and oblique relatives extracted from each corpus – 70% of both object relatives and oblique

relatives were of this type. This finding also makes clear that the type of object relative clause often used in experimental studies – namely those relative clauses containing full noun phrases that refer to animate entities (e.g. "*the senator that the reporter attacked*") – are rare in language experience, with only 1% for object relatives and 2% for oblique relatives. In this light and taking a constraint-satisfaction perspective, it is not surprising that they are difficult to comprehend (Gennari & MacDonald, 2008, 2009; Hsiao & MacDonald, 2016, 2013; Seidenberg & MacDonald, 2018).

(iii) ***Passive relative clauses***

Overall, passive relative clauses were less common than other types across all three corpora, except oblique relatives (Figure 2); they were, however, more common in texts written for older children and in nonfiction. There are several characteristics of passive relative clauses that inform why this might be. We first examined animacy. Previous studies have shown that when describing events that involve nouns of the same animacy or nouns that are conceptually similar, people tend to use alternative structures to circumvent competition (Gennari, Mirković, & Macdonald, 2012; Hsiao & MacDonald, 2016; Humphreys, Mirković, & Gennari, 2016). For example, an object relative clause that involves two similar animate nouns such as "***the girl*** *that* ***the woman*** *kissed*" is hard to produce. The close proximity of the two nouns (the agent and patient are separated by only a relative pronoun, or nothing for the reduced form) induces competition as the two animate nouns are equally good candidates for being the agent of the event. However, using a semantically equivalent passive relative clause to describe the same event, e.g. "***the girl*** *that was kissed by* ***the woman***", reduces competition as the two nouns are more distant and the planning of the agent can be deferred to the end position in the by-phrase. Competition is further alleviated if

the agent information is dropped entirely, as in "***the girl*** *that was kissed*". Comprehension mirrors the production patterns. Humphery et al. (2016) found that when the modified noun was animate, passive relative clauses were easier to comprehend than the semantically equivalent object relative clauses.

These production and comprehension constraints should be reflected in usage statistics, including in children's language. We therefore predicted that, even though there is a tendency for the patient or the object of an event to be inanimate, there would be more passive relative clauses with animate heads than object relative clauses with animate heads. To explore this, we extracted and hand-coded the animacy of the head noun in 100 passive relative clauses randomly sampled from each corpus. There was a tendency for the head noun to be inanimate, at 73% across corpus (77% for child-directed speech, 68% for picture books and 75% reading books), showing that the affectees of an event were more often inanimate. This mirrors the lexical-syntactic pattern seen in object relative clauses, although to a lesser extent, and confirms our hypothesis that passive relative clauses were used more often than object relatives to describe animate nouns. In summary, both passive and object relative clauses describe inanimate entities affected by an action usually performed by an animate agent; when there is a need to describe animate affectees, using the passive relative clause structure serves to distance the agent and patient, or to omit the agent entirely. This allows the speaker (or writer) to mitigate or avoid competition (Gennari et al., 2012).

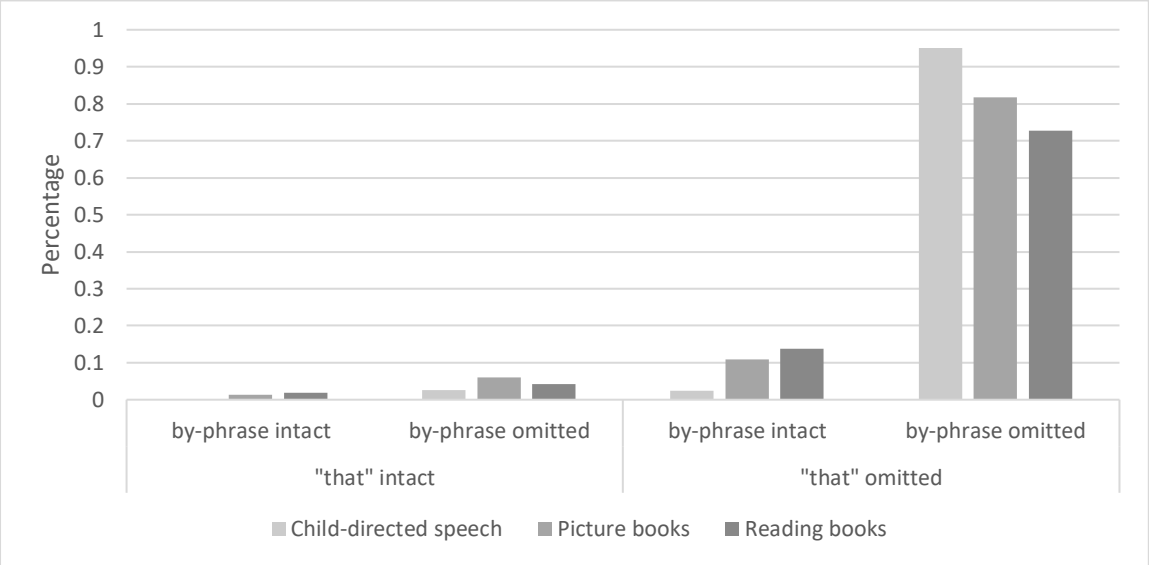
We then examined the rate of omission of the relative pronoun, such that a phrase like "*the girl that was kissed by the woman*" would be reduced to "*the girl kissed by the woman*". This pattern was common, with passive relative clauses being reduced across all three corpora, at 87% (97% in child-directed speech, 93% in picture books and 86% in reading books). Turning to the rate of by-phrase omission (e.g. "*the girl that was kissed*"), the agent

information ("*by the woman*") was more likely dropped overall, at 85% across all corpora (in child-directed speech at 98%, in picture books at 88% and in reading books at 84%).

Examining the distribution formed by crossing the omission of both the relative pronoun and the by-phrase, Figure 8 shows that most occurrences of passive relatives involved omission. Child-directed speech contained most omission and reading books the least. Overall, book language was more likely to preserve agent information through the by-phrase than spoken language.

Figure 8.

The percentage of passive relative clauses involving omission of relative pronoun and by phrase across corpus



General Discussion

Our findings extend and expand previous studies describing how the language in children's books is more syntactically complex than child-directed speech (Cameron-

Faulkner & Noble, 2013; Montag, 2019; Montag & MacDonald, 2015). Focussing on relative clauses, we built on the existing literature in several important ways. First, we examined language samples at scale by using large corpora and automated parsing procedures, complemented with hand-coding where appropriate. This allowed us to comprehensively capture the frequency and nature of relative clause usage within and across corpora. Second, we charted the development of relative clause usage in book language by comparing the content of picture books aimed primarily at pre-schoolers with books written for older children to read independently, and we compared each with child-directed speech. Furthermore, within the corpus of reading books we were able to take developmental slices by targeted age, and to examine genre by comparing fiction and nonfiction texts. Finally, we made links with the sentence processing literature by considering the frequency and distribution of key lexical-syntactic combinations within relative clause types and across corpora. Taken together, our findings show that book language is different to child-directed speech in terms of relative clause usage. This leads to the conclusion that by reading and listening to books, children experience sentence types that are rarely encountered in their spoken language environment. As experience is critical for learning, these findings have important implications for language acquisition.

We examined four types of relative clause: subject relatives, object relatives, oblique relatives, and passive relatives. All four types were more common in both types of book language than in child-directed speech. This finding resonates with previous research. Cameron-Faulkner and Noble (2013) analysed 20 picture books targeted at 2-year-olds. They found that there were more complete sentences (subject-predicate sentences) and complex sentences (sentences with more than one lexical verb) in the books than in sample of child-directed speech. Montag (2019) analysed a sample of 100 picture books and focused on

passive sentences (which we did not investigate here) and relative clauses. She found these complex sentences to be more frequent in picture books than in child-directed speech. Our study replicated Montag's findings for relative clauses: we found that all types were more frequent in book language than in speech. Additionally, within the two book corpora, relative clause usage increased with developmental level. This was demonstrated in two ways. First, there were more relative clauses in books written for independent reading than in picture books and second, within the reading corpus, there were clear increases in the frequency of relative clauses as the intended target age increased, as estimated by the Key Stage metadata. This was particularly evident for passive relative clauses. This suggests that exposure to rare structures is increasingly afforded by text, as its intended age and targeted reading level increases.

Across corpora, subject relative clauses were the most common type overall. Compared with other relative clause types within a corpus, however, we noted some important differences. Object relatives were most frequent in child-directed speech. By contrast, the picture book corpus and the reading corpus contained more subject relative clauses. Thus, while object relatives dominate child-directed speech (see also Montag, 2019), this is not the case in book language. This finding aligns with Roland et al.'s (2007) analysis of adult language. They found that object relatives dominated spoken language, as estimated by the British National Corpus Spoken portion and the Switchboard corpus, but were least frequent in written text, as estimated from sources such as the British National Corpus, the Brown Corpus and the Wall Street Journal Corpus, leading to the conclusion that object relative clauses are characteristic of more informal spoken language. This pattern was evident in our analyses of children's language, where the distribution of relative clause types in picture books already departed from child-directed speech in terms of reduced frequency of

object relative clauses compared to the high frequency of subject relatives. Our observation was not entirely in line with Montag (2019)'s finding that object relative clauses dominated in both picture books and child-directed speech. This may have been due to the fact that our picture corpus contained some chapter books (e.g. BFG) intended for both shared reading and independent reading. It seems likely that picture books as a category of book language is characterized by simpler but gradually more formal and bookish language, bridging the gap between speech and written text. This paves the way to increasing sophistication as written language develops, as evidenced by increases in the frequency of subject and passive relative in our reading book corpus, especially for nonfiction, as well as the adult texts analysed by Roland et al. (2007).

Within each type of relative clause, we also examined key lexical-syntactic features, including noun animacy and pronoun status. The processing literature tells us that adults find it easier to comprehend subject relative clauses that modify animate nouns. For object and oblique relative clauses, the pattern is opposite with comprehension being easier when the noun being modified is inanimate. These findings are well-replicated across several languages (English: Gennari & MacDonald, 2008, 2009; Mandarin Chinese: Hsiao & MacDonald, 2013; Wu, Kaiser, & Andersen, 2012; Dutch: Mak et al., 2002, 2006) and arise because when a transitive action is performed, animates are more likely to be the agent of an action whereas inanimate entities tend to be the affectee of an action. Children too are sensitive to animacy when producing or comprehending relative clauses (Arosio, Guasti, & Stucchi, 2011; Brandt et al., 2009; Kidd et al., 2007; Kirjavainen, Kidd, & Lieven, 2017; Lobo & Vaz, 2017). Consistent with this sensitivity in processing, these lexical-syntactic patterns were evident across all three corpora, indicating that they are apparent in children's language exposure.

The distributional patterns we saw in relative clause usage showed how animacy combines with other lexical features, such as pronoun use. For example, we found that object and oblique relatives were more likely to modify inanimate head nouns, and to contain an animate embedded relative clause noun, and that this was likely to be a pronoun, as in "*the book I read*" and "*the crayon he drew with*". These lexical-syntactic distributional features may serve discourse functions. Fox and Thompson (1990) argued that because pronouns represent given information, they serve to "anchor" the head noun in the discourse model. This allows new information expressed by the main clause to be linked to the already established information expressed by the relative clause. We observed that pronoun use in object and oblique relatives was higher in child-directed speech than in book language. This also suggests a possible difference in the need of anchoring or maintaining focus on the immediate interlocuters (e.g. *I, you*) within speech, compared with text. We also observed a high proportion of pronouns rather than full noun phrases, and relative pronouns (i.e. *which, that, who, whom*) tended to be omitted in object and oblique relatives. This observation highlights the rarity of certain types of relative clause in language input. It also aligns with Montag (2019) who found no instances of object relatives that modified animate nouns, contained full embedded noun phrase, and retained the relative pronoun at the same time, in children's picture book corpus or child-directed speech. Our findings also pattern with children's own speech, where object relative clauses are strongly skewed in favour of the animacy pairing of an inanimate head noun and an animate embedded noun, in addition to the embedding noun being a pronoun (most often first and second person pronouns) (Diessel, 2009). This pattern is also characteristic of adult language (Reali & Christiansen, 2007; Roland et al., 2007). Given this, it is not surprising that object relative clauses like "*the official that the reporter criticised*" are difficult to process and understand (Chomsky &

Miller, 1963; Hakes, Evans, & Brannon, 1976; Holmes & O'Regan, 1981; King & Just, 1991).

In summary, our work highlights the clear need to investigate differences between spoken and written language targeted at children. Books, even those written for pre-schoolers to hear in the context of shared reading, contain more relative clauses that child-directed speech. Our findings replicate and extend previous smaller scale studies (Cameron-Faulkner & Noble, 2013; Montag, 2019). They also complement parallel findings in the lexical domain showing that book language for young children is more rich and more diverse than day-to-day conversations (Dawson, Hsiao, Tan, Banerji, & Nation, 2021; Massaro, 2015; Montag, Jones, & Smith, 2015). As the sophistication of text grows with increases in targeted age, so too does the frequency of relative clause usage. Importantly, it is not just the number of relative clauses that changes but also their type and distribution: both picture books and reading books are dominated by subject relative clauses, different from speech, and book language contains dramatically more obliques and passives than child-directed speech. These changes are evident in the youngest developmental slice through the reading corpus, capturing books written for 5-7 year olds. There are also differences by genre with subject and passive relative clauses being more common in nonfiction.

Our study also demonstrates the merits of taking a corpus analysis approach to investigate large language samples. We recognise the limitations of current automated parsing procedures (e.g. some mis-identification of target structures by the parser, errors in search terms, difficulty in dealing with ill-structured text and unequal sample sizes). As computational advances continue, automated syntactic parsing should become more accurate. In the meantime, the levels of accuracy observed in this study are sufficient to draw meaningful conclusions that complement and extend findings from smaller-scale studies that

have relied on hand-coding (for broader discussion of strengths and limitations of different approaches, see Durrant, Brenchley, & McCallum, 2021).

Our findings make clear that once children can read and once they start to read widely, they will encounter language that is radically different from their day-to-day conversational experience. The corollary of this is that a lack of exposure to book language may limit children's language development. Given differences in syntactic complexity are apparent in picture books, this negative consequence may emerge early, if children are not engaged in shared reading. Book language provides unique access to grammar not easily encountered in speech. This has implications for the distributional lexical-syntactic features and associated discourse functions that children experience and from this, consequences for language development. Variability in access to book language is likely to emerge for many complex and interrelated reasons including lack of books in the home, reduced opportunity for shared reading, delays and difficulties in learning to read and low motivation to read. While the 'word gap' and negative sequalae in terms of vocabulary development are well-recognised, our findings highlight the pressing need to consider variations in exposure to complex grammar and the consequences of this for language and reading development.

References

Adani, F. (2011). Rethinking the acquisition of relative clauses in Italian: towards a grammatically based account. *Journal of Child Language*, *38*(1), 141–165. https://doi.org/DOI: 10.1017/S0305000909990250

Arosio, F., Guasti, M., & Stucchi, N. (2011). Disambiguating Information and Memory Resources in Children's Processing of Italian Relative Clauses. *Journal of Psycholinguistic Research*, *40*(2), 137–154. https://doi.org/10.1007/s10936-010-9160-0

Betancort, M., Carreiras, M., & Sturt, P. (2009). The processing of subject and object relative clauses in Spanish: An eye-tracking study. *Quarterly Journal of Experimental Psychology (2006)*, *62*, 1915–1929. https://doi.org/10.1080/17470210902866672

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge Approaches to Linguistics). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511804489

Booth, J. R., MacWhinney, B., & Harasaki, Y. (2000). Developmental differences in visual and auditory processing of complex sentences. *Child Development*, Vol. 71, pp. 981–1003. https://doi.org/10.1111/1467-8624.00203

Boyle, W., Lindell, A., & Kidd, E. (2013). Investigating the role of verbal working memory in young children's sentence comprehension. *Language Learning*, *63*, 211. https://doi.org/10.1111/lang.12003

Brandt, S., Kidd, E., Lieven, E., & Tomasello, M. (2009). The discourse bases of relativization: An investigation of young German and English-speaking children's comprehension of relative clauses. *Cognitive Linguistics, v.20, 539-570 (2009), 20*. https://doi.org/10.1515/COGL.2009.024

Cameron-Faulkner, T., & Noble, C. (2013). A comparison of book text and Child Directed Speech. *First Language*, *33*(3), 268–279. https://doi.org/10.1177/0142723713487613

Chomsky, N., & Miller, G. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (Vol. 2, pp. 269–320). New York, NY: John Wiley.

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, *2*. https://doi.org/10.1075/lab.2.3.01dab

Dąbrowska, E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences*, *28*, 604–615. https://doi.org/10.1016/j.langsci.2005.11.014

Dawson, N., Hsiao, Y., Banerji, N., Tan, A. W. M., & Nation, K. A. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*. https://doi.org/10.34842/5we1-yk94

Diessel, H. (2004). *The Acquisition of Complex Sentences (Cambridge Studies in Linguistics).* Cambridge: Cambridge University Press. *doi:10.1017/CBO9780511486531*

Diessel, H., & Tomasello, M. (2001). The Development of Relative Clauses in Spontaneous Child Speech. *Cognitive Linguistics*, *11*. https://doi.org/10.1515/cogl.2001.006

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, 81, 1–25.

Ding, D. D. (2002). The Passive Voice and Social Values in Science. *Journal of Technical Writing and Communication*, *32*(2), 137–154. https://doi.org/10.2190/EFMR-BJF3-CE41-84KK

Durrant, P., Brenchley, M., & McCallum, L. (2021). *Understanding Development and*

Proficiency in Writing: Quantitative Corpus Linguistic Approaches. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108770101

Fox, B., & Thompson, S. (1990). A Discourse Explanation of the Grammar of Relative Clauses in English Conversation. *Language*, *66*. https://doi.org/10.2307/414888

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, *58*(2), 161–187. https://doi.org/10.1016/j.jml.2007.07.004

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, *111*(1), 1–23. https://doi.org/10.1016/j.cognition.2008.12.006

Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, *65*(2), 141–176. https://doi.org/10.1016/j.cogpsych.2012.03.002

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*(6), 1411. https://doi.org/10.1037/0278-7393.27.6.1411

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, *51*(1), 97–114. https://doi.org/10.1016/j.jml.2004.02.003

Grodner, D., & Gibson, E. (2005). Consequences of the Serial Nature of Linguistic Input for Sentenial Complexity. *Cognitive Science*, *29*(2), 261–290. https://doi.org/10.1207/s15516709cog0000_7

Hakes, D. T., Evans, J. S., & Brannon, L. L. (1976). *Understanding sentences with relative*

clauses. *4*(3), 283–290.

Halliday, M. A. K. (1989). *Spoken and written language* (2nd ed). Oxford: Oxford University Press.

Holmes, V., & O'Regan, J. (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior*, *20*, 417–430. https://doi.org/10.1016/S0022-5371(81)90533-8

Hsiao, Y., & MacDonald, M. (2016). Production predicts comprehension: Animacy effects in Mandarin relative clause processing. *Journal of Memory and Language*, *89*. https://doi.org/10.1016/j.jml.2015.11.006

Hsiao, Y., & MacDonald, M. C. (2013). Experience and generalization in a connectionist model of Mandarin Chinese relative clause processing. *Frontiers in Psychology*, *4*, 767. https://doi.org/10.3389/fpsyg.2013.00767

Humphreys, G. F., Mirković, J., & Gennari, S. P. (2016). Similarity-based competition in relative clause production and comprehension. *Journal of Memory and Language*, *89*(C), 200–221. https://doi.org/10.1016/j.jml.2015.12.007

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343–365. https://doi.org/https://doi.org/10.1016/j.cogpsych.2010.08.002

Kidd, E., Brandt, S., Lieven, E., & Tomasello, M. (2007). Object relatives made easy : A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses children's processing of relative clauses. *Language and Cognitive Processes*, *22*(6), 860–897. https://doi.org/10.1080/01690960601155284

Kim, C. (2016). Processing Direct Object and Oblique Relative Clauses. *Language Research*, 52.2, 151-170.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of

working memory. *Journal of Memory and Language*, *30*(5), 580–602. https://doi.org/https://doi.org/10.1016/0749-596X(91)90027-H

Kirjavainen, Mi., Kidd, E., & Lieven, E. (2017). How do language-specific characteristics affect the acquisition of different relative clause types? Evidence from Finnish. *Journal Of Child Language*, *44*(1), pp120-157. https://doi.org/10.1017/S0305000915000768

Kitaev, N., & Klein, D. (2018). Constituency Parsing with a Self-Attentive Encoder. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics.

Kuhn, K. E., Rausch, C. M., Mccarty, T. G., Montgomery, S. E., & Rule, A. C. (2017). Utilizing Nonfiction Texts to Enhance Reading Comprehension and Vocabulary in Primary Grades. *Early Childhood Education Journal*, *45*(2), 285–296. https://doi.org/10.1007/s10643-015-0763-9

Lawrence, J. F. (2009). Summer Reading: Predicting Adolescent Word Learning from Aptitude, Time Spent Reading, and Text Type. *Reading Psychology*, *30*(5), 445–465. https://doi.org/10.1080/02702710802412008

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. *5th International Conference on Language Resources and Evaluation (LREC 2006).*

Lobo, M., & Vaz, S. (2017). Does the animacy of the antecedent play a role in the production of relative clauses? *Matraga*, *24*(41), 266–287. https://doi.org/10.12957/matraga.2017.28710

Logan, J. A. R., Justice, L. M., Yumus, M., & Chaparro-Moreno, L. J. (2019). When children are not read to at home: The million word gap. *Journal of Developmental & Behavioral Pediatrics*, *40*(5), 383–386.

MacDonald, M. C., Pearlmutter, N., & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, Vol. 101, pp. 676–703. https://doi.org/10.1037/0033-295X.101.4.676

Macdonald, R., Brandt, S., Theakston, A., & Lieven, E. (2020). The role of animacy in children's interpretation of relative clauses in English : Evidence from sentence–picture matching and eye movements. *Cognitive Science*. *44*, 1–35. https://doi.org/10.1111/cogs.12874

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of Memory and Language*, *47*, 50–68. https://doi.org/10.1006/jmla.2001.2837

Mak, W. M., Vonk, W., & Schriefers, H. (2006). Animacy in processing relative clauses: The hikers that rocks crush. *Journal of Memory and Language*, *54*(4), 466–490. https://doi.org/10.1016/j.jml.2006.01.001

Massaro, D. W. (2015). Two different communication genres and implications for vocabulary development and learning to read. *Journal of Literacy Research*, *47*(4), 505–527. https://doi.org/10.1177/1086296X15627528

Montag, J. L. (2019). Differences in sentence complexity in the text of children's picture books and child-directed speech. *First Language*, *39*(5), 527–546. https://doi.org/10.1177/0142723719849996

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The Words Children Hear: Picture Books and the Statistics for Language Learning. *Psychological Science*, *26*(9), 1489–1496. https://doi.org/10.1177/0956797615594361

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of

complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology. General*, *144*(2), 447.

Montgomery, J. W., & Evans, J. L. (2009). Complex sentence comprehension and working memory in children with specific language impairment. *Journal of Speech, Language, and Hearing Research: JSLHR*, *52*(2), 269–288. https://doi.org/10.1044/1092-4388(2008/07-0116)

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, Vol. 57, pp. 1–23. https://doi.org/10.1016/j.jml.2006.08.014

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379.

Rundblad, G. (2007). Impersonal, general, and social: The use of metonymy versus passive voice in medical discourse. *Written Communication*, *24*, 250–277. https://doi.org/10.1177/0741088307302946

Seidenberg, M. S., & MacDonald, M. C. (2018). The impact of language experience on language and reading: A statistical learning approach. *Topics In Language Disorders*, *38*(1), 66–83. https://doi.org/10.1097/TLD.0000000000000144

Tarone, E., Dwyer, S., Gillette, S., & Icke, V. (1998). On the use of the passive and active voice in astrophysics journal papers: With extensions to other languages and other fields. *English for Specific Purposes*, *17*(1), 113–132. https://doi.org/https://doi.org/10.1016/S0889-4906(97)00032-X

Traxler, M., Mason, R., Blozis, S., & Morris, R. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, *53*, 204–224. https://doi.org/10.1016/j.jml.2005.02.010

Traxler, M., Morris, R., & Seely, R. (2002). Processing subject and object relative clauses:

Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69–90. https://doi.org/10.1006/jmla.2001.2836

Wu, F., Kaiser, E., & Andersen, E. (2012). Animacy effects in Chinese relative clause processing. *Language and Cognitive Processes*, *27*(10), 1489–1524. https://doi.org/10.1080/01690965.2011.614423

Appendix A. List of books selected for the picture book corpus.

Title	**Author**
A dog with nice ears	Lauren Child
A Great Big Cuddle	Michael Rosen
A little bit Brave	Nicola Kinnear
A Squash and a Squeeze	Julia Donaldson
Aliens Love Underpants	Claire Freedman
All the colours I see	Allegra Agliardi
Along Came A Different	Tom McLaughlin
Animal Stories for 5 year olds	Helen Paiba
Barking for Bagels	Michael Rosen
Bedtime Stories for 5 year olds	Helen Paiba
Brown Bear, Brown Bear, What Do You See?	Bill Jnr Martin
But Excuse Me That is my Book	Lauren Child
Colin and Lee: Carrot and Pea	Morag Hood
Cyril and Pat	Emily Gravett
Dave the Lonely Monster	Anna Kemp
Dear Zoo	Rod Campbell
Dinosaur Roar!	Paul Stickland & Henrietta Stickland
Dogger	Shirley Hughes
Dogs don't do Ballet	Anna Kemp & Sara Ogilvie
Duck, Death, and the Tulip	Wolf Erlbruch
Each Peach Pear Plum	Allan Ahlberg & Janet Ahlberg
Elmer	David McKee
FArTHER	Grahame Baker-Smith
Fat Frog	Ruth Miskin
Five Minutes Peace	Jill Murphy
Fox & Goldfish	Nils Pieters
Fox's Socks	Julia Donaldson
Franklin's Flying Bookshop	Jen Campbell
Funny Stories for 5 Year Olds	Helen Paiba
George's Marvellous Medicine	Roald Dahl
Get up!	Ruth Miskin
Giraffe in the Bath and other tales	Russell Punter & Lesley Sims
Gracie la Roo goes to school	Marsha Qualey
Gracie la Roo sets sail	Marsha Qualey
Grandad's Island	Benji Davies
Granpa	John Burningham
Guess How Much I Love You	Sam McBratney
Hairy Maclary from Donaldson's Dairy	Lynley Dodd
Hampstead the Hamster	Michael Rosen
Heidi	Johanna Spyri
Hide and Seek	

Hide-and-Seek Pig	Julia Donaldson & Axel Scheffler
Hippo has a Hat	Julia Donaldson
Horrid Henry and the Secret Club	Francesca Simon
Horrid Henry tricks the Tooth Fairy	Francesca Simon
Horrid Henry: Ghosts and Ghouls	Francesca Simon
Horrid Henry's Halloween Horrors	Francesca Simon
How to be a Lion	Ed Vere
Hubert Horatio How to raise your grown-ups	Lauren Child
I can hop	Ruth Miskin
I Need a New Bum	Dawn McMillan
I Want My Hat Back	Jon Klassen
If all the world were...	Joseph Coelho & Allison Colpoys
In the Bath	Ruth Miskin
Into the Forest	Anthony Browne
Is it a Mermaid?	Candy Gourlay
John Brown, Rose and the Midnight Cat	Jenny Wagner
Joy	Corrinne Averiss
Kitchen Disco	Clare Foges & Al Murphy
Little Beauty	Anthony Browne
Looking for Atlantis	Colin Thompson
Lost and Found	Oliver Jeffers
Loved To Bits	Teresa Heapy & Katie Cleminson
Magical Stories for 5 year olds	Helen Paiba
Me and my Fear	Francesca Sanna
Michael Rosen's Sad Book	Michael Rosen
Mog the Forgetful Cat	Judith Kerr
Monkey Puzzle	Julia Donaldson
Mr Men: Chinese New Year	Adam Hargreaves
Murray the Race Horse	Gavin Puckett
My Father's Arms are a Boat	Stein Erik Lunde
Nice Work for the Cat and the King	Nick Sharratt
Night-Time Cat	Julia Tedd
Nip and Chip	Ruth Miskin
No-Bot	Sue Hendra & Paul Linnet
Nog in the Fog	Ruth Miskin
Odd Dog Out	Rob Biddulph
of Thee I sing	Barack Obama
Oi Cat!	Kes Gray
Oi Dog!	Kes & Claire Gray
Oi Frog!	Kes Gray
Oi Goat!	Kes Gray
Owl Babies	Martin Waddell & Patrick Benson
Pants	Giles Andreae

Peace at Last	Jill Murphy
Peck peck peck	Lucy Cousins
Peppa goes to London	Lauren Holowaty
Peppa meets Father Christmas	Lauren Holowaty
Peppa the Mermaid	Lauren Holowaty
Peppa's Magical Unicorn	Lauren Holowaty
Princess Mirror-Belle and the Flying Horse	Julia Donaldson
Princess Mirror-Belle and the Sea Monster's Cave	Julia Donaldson
Rabbit & Bear Attack of the Snack	Julian Gough
Rabbit & Bear The Pest in the Nest	Julian Gough
Rabbityness	Jo Empson
Raccoon on the Moon	Russell Punter
Rag the Rat	Ruth Miskin
Red Ned	Ruth Miskin
Room on the Broom	Julia Donaldson
Rosie's Walk	Pat Hutchins
Ruby Red Shoes Goes to London	Kate Knapp
Ruby's Worry	Tom Percival
Run, run, run!	Ruth Miskin
Sharing a Shell	Julia Donaldson
Sophie Johnson Unicorn Expert	Morag Hood
Squishy McFluff the Invisible Cat: Seaside Rescue!	Pip Jones
Stardust	Jeanne Willis
Stick Man	Julia Donaldson
Sun Hat Fun	Ruth Miskin
Superworm	Julia Donaldson
Sweep	Louise Greig & Julia Sarda
That's not my puppy...	Fiona Watt
That's Not my Unicorn…	Fiona Watt
The Bad-Tempered Ladybird	Eric Carle
The BFG	Roald Dahl
The Building Boy	Ross Montgomery
The Bumblebear	Nadia Shireen
The Cat in the Hat	Dr Seuss
The Day the Crayons Quit	Drew Daywalt & Oliver Jeffers
The Day War Came	Nicola Davies
The Detective Dog	Julia Donaldson
The Flat Rabbit	Bardur Oskarsson
The Gift	Carol Ann Duffy
The Gruffalo	Julia Donaldson
The Gruffalo's Child	Julia Donaldson
The Heart and the Bottle	Oliver Jeffers
The Highway Rat	Julia Donaldson
The Jolly Christmas Postman	Janet Ahlberg & Allan Ahlberg

The Jolly Postman or Other People's Letters	Janet Ahlberg & Allan Ahlberg
The Last Chip: The Story of a Very Hungry Pigeon	Duncan Beedie
The Lion Inside	Rachel Bright
The Marvellous Moon Map	Teresa Heapy & David Litchfield
The Memory Tree	Britta Teckentrup
The Owl who was Afraid of the Dark	Jill Tomlinson
The Paper Dolls	Julia Donaldson
The Pond	Nicola Davies
The Scar	Charlotte Moundlic
The Smartest Giant in Town	Julia Donaldson
The Snail and the Whale	Julia Donaldson
The Storm Whale	Benji Davies
The Storm Whale in Winter	Benji Davies
The Tiger Who Came to Tea	Judith Kerr
The Twits	Roald Dahl
The Ugly Five	Julia Donaldson
The Very Hungry Caterpillar	Eric Carle
The Wonky Donkey	Craig Smith
Tiddler	Julia Donaldson
Tug, tug	Ruth Miskin
Very little Cinderella	Teresa Heapy & Sue Heap
We're Going on a Bear Hunt	Michael Rosen
What Happens Next	Shinsuke Yoshitake
What is Poo?	Katie Daynes
Whatever Next!	Jill Murphy
When Sadness comes to call	Eva Eland
Where the Wild Things Are	Maurice Sendak
Where's Spot?	Eric Hill
Willy and the Cloud	Anthony Browne
Willy the Wimp	Anthony Browne
Witchfairy	Brigitte Minne
Zog	Julia Donaldson
Zog and the Flying Doctors	Julia Donaldson

Appendix B. List of CHILDES corpora included in the child-directed speech corpus

Corpus	Child age range	n	Reference
Belfast	2;0-4;5	8	Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. New York: Oxford University Press.
Gathercole/Burns	3;0-6;4	12	Gathercole, V. (1986). The acquisition of the present perfect: explaining differences in the speech of Scottish and American children. *Journal of Child Language*, *13*, 537–560
Howe	1;6-1;8 (session 1) 1;11-2;1 (session 2)	16	Howe, C. (1981). *Acquiring language in a conversational context.* New York: Academic Press.
Korman	6-16 weeks	6	Korman, M., & Lewis, C. (2001). Mothers' and fathers' speech to their infants: Explorations of the complexities of context. In M. Almgren, A. Barreña, M.-J. Ezeizabarrena, I. Idiazaabal, & B. MacWhinney (Eds.), *Research on child language acquisition* (pp. 431-453). Somerville, MA: Cascadilla Press
Lara	1;9-3;3	1	Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1-21. doi:10.1016/j.cogpsych.2017.07.002.
Manchester	1;8-3;0	12	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language, 28*, 127-152.
MPI-EVA Manchester	1;8-3;2	4	Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20* (3), 481-508.
Nuffield	0;11	76	McGillion, M., Pine, J. M., Herbert, J. S., & Matthews, D. (2017). A randomised controlled trial to test the effect of promoting caregiver contingent talk on language development in infants from diverse socioeconomic status backgrounds. *Journal*

			of Child Psychology and Psychiatry, 58 (10), 1122-1131
Tommerdahl	2;6-3;6	23	Tommerdahl, J. and Kilpatrick, C. (2014). The Reliability of Morphological Analyses in Language Samples. *Journal of Language Testing, 31* (1), 3-18.
Wells	1;6-5;0	32	Wells, C. G. (1981). *Learning through interaction: The study of language development.* Cambridge, UK: Cambridge University Press.

Appendix C. Tregex search pattern for different types of relative clauses

Relative clause type	Tregex pattern	Example
Subject relative clause	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(VBD|VBG|VBZ|VB|VBP|<'s'|'re|'s|are|was|were|be)<NP))))	the man who kicked the ball
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<CC<<((VBD|VBG|VBZ|VB|VBP|<'s'|re|'s|are|was|were|be)$NP))))	the man who can kick the ball and win the game
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP!<S<((VBD|VBG|VBZ|VB|VBP<have|has|had$(VP<NP)))))))	the man who had kicked the ball
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<MD<(VP!<(VB<be)<NP))))	the man who can kick the ball
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(VBD|VBG|VBZ|VB|VBP<'s'|re|'s|are|was|were)<(VP<VBG<NP))))))	the man who is kicking the ball
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<MD<(VP<(VB<be)<(VP<VBG<NP))))))	the man who could be kicking the ball
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(VBD|VBG|VBZ|VB|VBP<'s'|re|'s|are|was|were|be)<NP))))	the man who is a football player
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(VBD<had|have|has$(VP<(VBN<been)<NP))))))	the man who had been a football player
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(MD$,VP<(VB<have$(VP<NP)))))))	the man who could be a football player
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(VBD|VBG|VBZ|VB|VBP|MD<had|have|has|would|could|should|will|can|may|might$(VP<(VP<(NP|ADJP))))))))	the man who could have kicked the ball
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(VBD|VBG|VBZ|VB|VBP|MD<had|have|has|would|could|should|will|can|may|might$(VP<(VP<(NP|ADJP))))))))	the man who had been kicking the ball
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP!<(VP<(VBN|VBD)!<S+(VBD|VBG|VBZ|VB|VBP!<'s|s|'re|are|was|were|be|$NP))))	the man who jumped
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<CC<<((VBD|VBG|VBZ|VB|VBP!<'s|'re|s|are|was|were|be)|$NP))))	the man who jumped and kicked
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<MD<(VP!<(VB<have)!<S!<(VB<be)<NP))))	the man who can jump
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(VBD|VBG|VBZ|VB|VBP<'s'|re|'s|are|was|were)<(VP<VBG!<NP))))	the man who is jumping
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<MD<(VP<(VB<be)<(VP<VBG!<NP))))	the man who can be jumping
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whatever|how|however)<(S<(VP<(VBD|VBG|VBZ|VB|VBP<'s'|re|'s|are|was|were)!<NP<(ADJP|PP)))))	the man who is happy
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whatever|how|however)<(S<(VP<(VBD|VBG|VBZ|VB|VBP<had|have|has$(VP<(VBN<been)!<NP<(ADJP|PP)))))	the man who has been happy
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whatever|how|however)<(S<(VP<(VBD<had|have|has$(VP<(VBN<been)!<NP<(ADJP|PP)))))))	the man who has been jumping
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(MD$,VP<(VB<be!$NP$(ADJP|PP)))))))	the man who can be happy
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(MD$,VP<(VB<be!$NP$(VP!<NP<VBG))))))))	the man who could be jumping
	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whatever|how|however)<(S<(VP<(MD$,VP<(VB<have$(VP!<S<(VBN!<been)!<NP))))))))	the man who could have jumped
	NP|VP<(NP$++(SBAR<WHNP<(S<(VP<(VBD|VBG|VBZ|VB|VBP|MD<had|have|has|would|could|should|will|can|may|might$(VP!<(VP|VBN<be|been)<(VP!<VP!<NP!<S))))))))	the man who had been jumping
Object relative clause	NP|VP<(NP$++(SBAR!<WHPP<((WHNP!<<what|whose|wohever|whatever|how|however)<(S<NP<(VP!<NP!<<(PP<;IN|RP|TO)!<S)))) NP!,IN<(SBAR<(S<NP<(VP!<NP!<<(PP<;IN|RP|TO)!<S)))	the fun that I had the fun I had
Oblique relative clause	NP<(SBAR<(WHNP!<<what|whose|whatever|how|however)<(S<NP<(VP!<NP!<<(PP<;IN|RP|TO)!<S))) NP|VP<(NP$++(SBAR<((WHPP<WHNP!<<what|whose|whoever|whatever|how|however)<(S<NP<(VP!<NP!<<(PP<;IN|RP|TO)!<S))) NP!,IN<(SBAR<(S<NP<(VP!<NP<<(PP<;IN|RP|TO)!<S)))	the crayon that you draw with the crayon with which I used the crayon you drew with
Passive relative clause	NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S<(VP<(VBD<is|are|was|were|be|get|s|get|go)<(VP<VBN!<(PP<(IN<by))))))) NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whatever|how|however)<(S<(VP<(MD$,(VP<(VB<be$(VP<(VBN!$(PP<(IN<by))))))))) NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S!<NP<((VP<(VBD|VBZ|VB|VBP|MD<had|have|has|would|could|should|will|can|may|might)$(VP<(VBN<been$(VP<VBN!<(PP<(IN<by))))))))) NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whatever|how|however)<(S<(VP<(VBD|VBZ|VB|VBP|MD<had|have|has|would|could|should|will|can|may|might)$(VP<(VBN<be|been$(VP<VBN<(PP<(IN<by))))))))) NP|VP<(NP$++(SBAR<(WHNP|<<what|whose|whoever|whatever|how|however)<(S!<NP<(VP<(MD$(VP<(VBD|VBZ|VB|VBP|MD$(VP<(VBN<be|been$(VP<VBN<(PP<(IN<by))))))))) NP<NP<(VP<VBN!,(PP<(IN<by)))) NP,(VP<(VBN$,.(PP<(IN<by))))	the houses that were built the houses that can be built the houses that have been built the houses that could have been built the houses that were built by the workers the houses that can be built by the workers the houses that have been built by the workers the houses that might have been built by the workers the houses built the houses built by the workers

Appendix D. Error analysis on the false positives and false negatives that machine extraction method produced among the 1000 random sentences extracted from each corpus. "NA" under Error source indicate those adverbial clauses that were indistinguishable from the object relative clause structure.

Corpus	Relative clause type identified	Sentence (relative clause italicised)	Error analysis	Error source	Error type
Child-directed speech	Object	"*Oh sorry the longest the longest it is.*"	Speech with repetitions	Text	False positive
	Object	"*Is that all I get*?"	"that" mistaken as preposition by parser	Parser	False negative
Picture books	Object	"*That's the way they are.*"	Adverbial clause	NA	False positive
	Object	"*How will Dad know how to make my toast the way I like it*."	Adverbial clause	NA	False positive
	Object	"*Another present Rosemary would have liked was a boat*."	"present" was parsed as an adjective	Parser	False negative
	Object	"*I just love the way you talk*."	Adverbial clause	NA	False positive
	Object	"*I'm wearing the black patent leather shoes with the blue flowers I always wear*."	"wear" was parsed as a punctuation	Parser	False negative
	Oblique	"*There wasn't much Hubert didn't excel at*."	"much Herbert" was parsed as a single NP	Parser	False negative
	Passive	"*A book shop… a shoe shop… a florist… a tailor's a toy store… a hairdressing salon… but HOLD ON!*"	"hold" was parsed as a past participle	Parser	False positive
Reading books	Object	"What manner of landlord could this be , who made a point of knowing his tenants as men and women the	Adverbial clause	NA	False positive

	moment he came to the estate ?"			
Object	"*Communication can occur through gestures , eye contact (or the lack of it) , touch , and even the way you stand when you look at someone* ."	Adverbial	NA	False positive
Object	"*It contained all my worldly possessions : some clothes wrapped round the book grandmere used to teach me to read , a spoon and a knife*."	"the book grandmere" parsed as a single NP	Parser	False negative
Oblique	"*I 'll tell you something that Mr. Jack Rabbit told about.*"	"told" was parsed as a verb of past tense but not dominated by a VP	Parser	False negative

Appendix E. Lexical syntactic distribution of subject relative clauses

Table 1. Frequency of transitive and instransitive subject relative clauses per 1000 noun phrases (raw frequency in parenthesis) across the three corpora

	Child-directed speech	Picture books	Reading books
transitive	0.48 (695)	2.25 (249)	4.32 (36172)
intransitive	0.69 (1000)	2.97 (329)	4.84 (40586)

Table 2. Frequency per 1000 noun phrases (raw frequency in parenthesis) and percentage of subject relative clauses containing "be" verbs across the three corpora

	Child-directed speech	Picture books	Reading books
Frequency	0.40 (583)	0.99 (110)	1.37 (11469)
copula%	34%	19%	15%

Table 3. Frequency of transitive and instransitive subject relative clauses per 1000 noun phrases (raw frequency in parenthesis) in fiction and nonfiction in the reading book corpus

	Fiction	Nonfiction
transitive	3.95 (26454)	5.63 (9431)
intransitive	4.86 (32593)	4.71 (7881)

Table 4. Frequency per 1000 noun phrases (raw frequency in parenthesis) and percentage of subject relative clauses containing "be" verbs in fiction and nonfiction in the reading book corpus

	Fiction	Nonfiction
Frequency	1.29 (8614)	1.61 (2687)
copula%	15%	16%

Table 5. Percentages of animate head nouns and embedded nouns in the 100 randomly sampled transitive subject relative clauses from each of the three corpora

	Child-directed speech	Picture books	Reading books
Animate head nouns	60%	69%	44%
Animate embedded nouns	83%	64%	69%

Table 5. Percentages of animate head nouns in the 100 randomly sampled intransitive subject relative clauses from each of the three corpora

	Child-directed speech	Picture books	Reading books
Animate head nouns	34%	41%	39%

Appendix F. Lexical syntactic distribution of object and oblique relative clauses

Table 1. Relative pronoun omission in object and oblique relative clauses by percentage (and by frequency per 1000 noun phrases and raw counts in the parenthesis) in the three corpora

Relative pronoun omission	Child-directed speech	Picture books	Reading books
object relative clauses	77% (1.48, 2148)	77% (3.08, 341)	64% (3.14, 26326)
oblique relative clauses	56% (0.11, 156)	68% (0.31, 34)	31% (0.35, 2972)

Table 2. Percentages of animate head nouns in the 100 randomly sampled object relative clauses and 100 oblique relative clauses from each of the three corpora. Note that there were only 42 oblique relative clauses in the picture book corpus.

Inanimate head nouns	Child-directed speech	Picture books	Reading books
object relative clauses	98%	97%	94%
oblique relative clauses	93%	88%	90%

Table 3. Percentages of the embedded nouns being pronouns in the 100 randomly sampled object relative clauses and 100 oblique relative clauses from each of the three corpora. Note that there were only 42 oblique relative clauses in the picture book corpus.

Embedded noun being pronoun	Child-directed speech	Picture books	Reading books
object relative clauses	91%	65%	65%
oblique relative clauses	84%	69%	74%

Table 4. Percentages of animate embedded nouns in the 100 randomly sampled object relative clauses and 100 oblique relative clauses from each of the three corpora. Note that there were only 42 oblique relative clauses in the picture book corpus.

Animate embedded noun		Child-directed speech	Picture books	Reading books
object relative clauses	pronoun	97%	100%	100%
	Full NP	89%	77%	86%
oblique relative clauses	Pronoun	99%	100%	100%
	Full NP	69%	92%	81%

Appendix G. Lexical syntactic distribution of passive relative clauses

Table 1. Percentages of inanimate head nouns and embedded nouns in the 100 randomly sampled passive relative clauses from each of the three corpora

	Child-directed speech	**Picture books**	**Reading books**
Inanimate head nouns	77%	68%	75%

Table 2. Percentages of relative pronoun omission in the 100 randomly sampled passive relative clauses from each of the three corpora

	Child-directed speech	**Picture books**	**Reading books**
Relative pronoun omission	97%	93%	86%

Table 3. Percentages of by-phrase omission in the 100 randomly passive relative clauses from each of the three corpora

	Child-directed speech	**Picture books**	**Reading books**
By-phrase omission	98%	88%	84%

Table 4. Percentages of inanimate head noun, relative pronoun omission, and by-phrase omission in the 100 randomly passive relative clauses from each of the three corpora

			Child-directed speech	**Picture books**	**Reading books**
Animate head noun	Relative pronoun intact	By-phrase intact	0%	1%	9%
		By-phrase omitted	1%	3%	8%
	Relative pronoun omitted	By-phrase intact	0%	4%	1%
		By-phrase omitted	22%	24%	7%
Inanimate head noun	Relative pronoun intact	By-phrase intact	0%	0%	11%
		By-phrase omitted	13%	5%	19%
	Relative pronoun omitted	By-phrase intact	0%	10%	23%
		By-phrase omitted	64%	53%	22%