

## Putting lexical and semantic cues into discourse context: A multifactorial corpus study of relative clauses in child-directed and child speech

Kin Chung Jacky Chan  
Lancaster University, UK  
Durham University, UK

Anna Theakston  
University of Manchester, UK

Silke Brandt  
Lancaster University, UK

**Abstract:** Object relative clauses (RCs) are, in general, more difficult to process than subject RCs. Yet, if the noun phrases (NPs) in an object RC meet certain lexical, semantic, and discourse-based expectations, this processing asymmetry disappears. However, the effect of discourse-based expectations has only been studied in adults. This is the first study to examine the nature of discourse-based features of NPs, alongside lexical and semantic NP features, in complex sentences in naturalistic child-directed and child speech through English developmental (2-5 years) corpora. We analyzed 1126 and 495 RCs from caregivers and children respectively. We found that the lexical and semantic characteristics of the NPs in subject and object RCs are similar to those previously observed in adult-directed speech and do not show great differences between child and child-directed speech. For the discourse-based features, however, we found differences between previously reported studies of adult-directed speech and our analyses of child and child-directed speech. Whereas in adult-directed speech, object RCs tend to have discourse-new head NPs, most of the object RCs in child and child-directed speech have discourse-old head NPs. In addition, both children and caregivers tend to use non-topical NPs within object RCs. We conclude that the distribution of discourse-based information in child and child-directed speech is different from that in adult-directed speech. We further discuss how these distributional patterns shape the expectations that language users bring to comprehension and production.

**Keywords:** lexical characteristics; discourse information; relative clause; complex sentence; corpus analysis

**Corresponding author(s):** Kin Chung Jacky Chan, Department of Psychology, Durham University, Upper Mountjoy, South Road, Durham, DH1 3LE, UK. Email: [kin.c.chan@durham.ac.uk](mailto:kin.c.chan@durham.ac.uk).

**ORCID ID(s):** Kin Chung Jacky Chan: <https://orcid.org/0000-0002-8533-2949>; Anna Theakston: <https://orcid.org/0000-0002-9483-7893>; Silke Brandt: <https://orcid.org/0000-0003-3363-8740>

**Citation:** Chan, K. C. J., Theakston, A., & Brandt, S. (2026). Putting lexical and semantic cues into discourse context: A multifactorial corpus study of relative clauses in child-directed and child speech. *Language Development Research*, 6(1), 345–407. <http://doi.org/10.34842/ldr2026-998>

## Introduction

Frequency effects have been shown to play a role in most, if not all, areas of language acquisition and processing in both children and adults (for overviews see Ambridge et al., 2015, Diessel, 2007; Ellis, 2002; Pfänder & Behrens, 2016). As pointed out by Ambridge et al. (2015, p. 241), frequency effects can occur at different levels, ranging from “concrete lexical strings” to “abstract categories (e.g., particular orderings of SUBJECT and OBJECT) and cues (e.g. animacy, givenness)”. However, in language acquisition research, some levels and cues have received less attention than others. Compared to lexical and syntactic cues (e.g., mapping between SUBJECT and specific lexical forms or word order), discourse-based cues (e.g., mapping between SUBJECT and givenness or topichood) are currently under-investigated (for some exceptions see Song & Fisher, 2005; Theakston, 2012). In the current study, we investigate a wide range of cues within subject and object relative clauses (RCs). In particular, whereas previous studies on RC acquisition have primarily focused on semantic and lexical cues (e.g., Brandt et al., 2009; Kidd et al., 2007), we will present a more comprehensive picture by also investigating givenness and topichood of SUBJECT and OBJECT referents and related cues such as definiteness. Our investigation is based on spontaneous child and child-directed speech. If we find that subject and object RCs produced by children and their caregivers contain specific discourse-based cues, this will allow us and other language acquisition researchers to derive hypotheses for how children might use these cues together with semantic and lexical cues when processing RCs.

According to experience- and usage-based accounts, language processing is guided by the distribution of lexical, semantic, and discourse-based features in language production (e.g., Ambridge et al., 2015; MacDonald, 2013; Reali & Christiansen, 2007). This hypothesis has been extensively tested on children’s and adults’ processing and comprehension of RCs, particularly object RCs. In spontaneous speech, object RCs are overwhelmingly used with inanimate head NPs (e.g., *your new skirt* in (1)) and pronominal embedded NPs (e.g., *you* in (1)).

- (1) **Your new skirt** that **you** got for your birthday. [Gina’s mum, 3;00.50]

When these semantic and lexical constraints are met, children and adults find it relatively easy to process and comprehend object RCs in an experimental context (e.g., Brandt et al., 2009; Kidd et al., 2007; Gennari & MacDonald, 2008; MacDonald et al., 2020). However, when these constraints are violated, children and adults find it more difficult to process and comprehend object RCs (2) than subject RCs (3).

(2) **The dog** [CP that [IP **the cat** [VP was chasing \_\_ ]]] was fast.<sup>1</sup>

(3) **The dog** [CP that [IP \_\_ [VP was chasing **the cat**]]] was fast.

For adults, it has also been shown that their RC processing is driven by discourse-level characteristics of the head and embedded NPs. For example, based on distributional patterns in language production, adults expect the head NPs of object RCs to be discourse-new and the embedded NPs of object RCs to be discourse-old (e.g., Fox & Thompson, 1990; Roland et al., 2012). Whether, based on their input and own RC production, children can be expected to show sensitivity to the same discourse-level characteristics of subject and object RCs is currently unknown and will be explored in the current study. Before we present and discuss our comprehensive analysis of lexical, semantic, and discourse-based cues within RCs produced by children and their caregivers, we will summarize previous studies and alternative theoretical approaches of RC acquisition and processing and present our research questions and hypotheses in more detail.

### **The Relative Clause Processing Asymmetry**

In the absence of any lexical, semantic, or discourse-based constraints, both children and adults are generally faster or better at processing subject RCs (3) compared to object RCs (2) (e.g., de Villiers et al., 1979; Frazier & Clifton, 1989; Friedmann & Novogrodsky, 2004; Gordon et al., 2001; Tavakolian, 1981). This asymmetry has been explained in terms of object RCs being syntactically more complex (e.g., Friedmann & Novogrodsky, 2004; Miyamoto & Nakamura, 2003) or cognitively more demanding to process (e.g., Gibson, 1998).

Object RCs are syntactically more complex than subject RCs in two ways. First, as stated by the Structural Distance Hypothesis (O'Grady et al., 2003), the number of syntactic tree nodes intervening between the head NP (*the dog* in (2) and (3)) and the gap (indicated by underscores in (2) and (3)) is greater for object RCs than for subject RCs. In a subject RC (3), the head NP is extracted from the gap that is within the inflectional phrase (IP), whereas in an object RC (2), the head NP is extracted from the gap within the verb phrase (VP), which is embedded within the IP. Therefore, the syntactic distance between the head NP and the gap in object RCs is longer than that in subject RCs. The other source of complexity is based on Rizzi's (1990) theory of Relativized Minimality. This theory suggests that if the head NP and the gap is intervened by another NP, as is the case for object RCs, the dependency between the head NP and the gap is more difficult to establish (Friedmann et al., 2009). This is especially true if the intervening NP is of the same form (e.g., the head and embedded

---

<sup>1</sup> CP stands for complementizer phrase; IP stands for inflectional phrase; and VP stands for verb phrase.

NPs are both lexical NPs or pronouns; Gordon et al., 2001).

In terms of the demand on processing capacity, according to the Dependency Locality Theory of sentence comprehension (Gibson, 1998, 2000; Warren & Gibson, 2002), the dependency of two elements is more demanding if they are further apart from each other. The linear distance between head NPs and gaps are greater in object compared to subject RCs. As a result, a language user has to maintain the head NP in memory for a longer period of time and over more constituents before it can be integrated with the post-verbal gap when processing object RCs.

It has also been suggested that the difficulty associated with processing object RCs is due to them not following canonical word order, that is the order of agent and patient in object RCs does not resemble that found in simple transitive sentences (e.g., Bever, 1970; Diessel & Tomasello, 2005). For instance, in languages with canonical SVO word order and head-initial RCs (where the RC follows the head NP), such as English and German, a language user can arrive at the correct interpretation of subject RCs when they assign agent and patient roles to NPs as they would for a simple transitive sentence. However, this parse fails in the processing of object RCs where the agent-patient order shows a reverse pattern than that in simple transitive sentences. When this happens, syntactic reanalysis is needed for the language user to correctly interpret the RC.

### **Experience-Based Processing of Object Relative Clauses**

In contrast to the syntactic complexity and memory-based approaches discussed in the previous section, experience- and usage-based approaches suggest that object RCs have been found to be more difficult to process because the test sentences used in most previous experimental studies violate expectations for the form and function of object RCs. In support of experience- and usage-based approaches, a series of studies with adults and children between the age of 3 and 5 years (e.g., Brandt et al., 2009; Fox & Thompson, 1990; Gennari & MacDonald, 2008; Kidd et al., 2007; MacDonald et al., 2020; Traxler et al., 2002) have demonstrated that the difficulty associated with the processing of object RCs can be eliminated or reduced if the head NP is inanimate (see examples (4b), (4c), and (4d)). Studies have also found that object RCs are easier to process if the embedded NP is a pronoun, especially a (first- or second-person) personal pronoun (Brandt et al., 2016; Fox & Thompson, 1990; Haendler et al., 2015; Kidd et al., 2007; Warren & Gibson, 2002; see examples (4c) and (4d)). See (4a) to (4d) below in decreasing order of processing difficulty. These findings have been attributed to the fact that, in spontaneous speech, objects tend to be inanimate (e.g., Mak et al., 2002, 2006) and subjects tend to be expressed by pronouns (Du Bois, 1987). As mentioned above, most object RCs produced and encountered by children and adults also display these semantic and lexical cues (e.g., Diessel, 2009; Fox & Thompson, 1990; Kidd et al., 2007).

- (4a) The dog that the man was carrying was heavy.
- (4b) The bag that the man was carrying was heavy.
- (4c) The bag that he was carrying was heavy.
- (4d) The bag that I was carrying was heavy.

### Discourse-Based Cues

The accounts discussed thus far suggest that syntactic complexity, memory, or the lack of lexical and semantic cues are plausible explanations for the difficulty associated with processing object RCs in previous experimental studies. However, they are heavily based on experiments that examined the processing of isolated sentences. In natural language, sentences are rarely produced in isolation, and language users also generate expectations based on how words and structures are used in discourse. In an analysis of adult-directed speech, Fox and Thompson (1990) investigated the form and function of subject and object RCs and provided a discourse-based explanation of the formal patterns of the head and embedded NPs. They analyzed 414 RCs in adult spoken English and examined the lexical, semantic, and discourse-based characteristics of the head and embedded NPs in subject and object RCs. Object RCs (e.g., *the car that she borrowed had a low tire* (Fox & Thompson, 1990, p. 303)) usually use discourse-old referents (e.g., *she*) to anchor discourse-new referents (e.g., *the car*) introduced in the main clause. These discourse-new referents also tend to be inanimate and non-human. This leads to object RCs often being used with inanimate head NPs and pronominal embedded NPs. In contrast, the embedded NPs of subject RCs usually contain new information, as subject RCs mostly characterize the head NP that is previously mentioned in the discourse context (e.g., *You should have, it was **his** birthday. **Who** was it? **That little boy** who gave you the sweets* (Gina's mum, 3;00.19)). This results in the embedded NP of subject RCs being more likely to be an indefinite lexical NP (e.g., *there is a woman in my class who is **a nurse*** (Fox & Thompson, 1990, p.301)). Owing to the different discourse functions of subject and object RCs – anchoring for object RCs and characterizing for subject RCs – Fox and Thompson (1990) also found that subject RCs tend to be used with indefinite, rather than definite, lexical head NPs (e.g., *he's got **a spring** that comes way up* (Fox & Thompson, 1990, p.301)).

The patterns described by Fox and Thompson (1990) suggest that the processing of RCs might also be driven by discourse-based cues and expectations, as has been shown for lexical and semantic cues (see previous section). In fact, in a series of questionnaires and self-paced reading tasks with English-speaking adults, Warren and Gibson (2002) found that the ease of processing of object RCs is dependent on the referential complexity of the embedded NP, which is usually linked to discourse-based features (e.g., givenness). In line with the Givenness Hierarchy (Gundel et al., 1993), object RCs with a pronoun as the embedded NP were easier and quicker to

process than those with a proper noun, followed by those with a definite NP; those with an indefinite NP were the slowest and most difficult to process. These results seem to suggest that givenness of the embedded NP plays a role in the processing of object RCs, as given referents are more likely to be expressed by pronouns, proper nouns, or definite NPs. However, Warren and Gibson's (2002) manipulation of the form of embedded NPs cannot be viewed as a direct manipulation of givenness, as the test sentences were still presented in isolation and without any context. This means that they looked at lexical features (i.e., definiteness of NPs) and assumed that they can be linked to discourse-based features (i.e., givenness), without directly testing the role of discourse-based features. Nonetheless, their results have demonstrated that, in line with other studies (e.g., Fox & Thompson, 1990; Kidd et al., 2007), embedded NP type (whether an embedded NP is a lexical NP or a pronoun) and definiteness of the embedded NP play a role in the processing of object RCs.

Another discourse-based characteristic that has been looked at in relation to RC processing is topichood (i.e., whether an entity is the focus of the ongoing discourse). Mak et al. (2008) postulated the Topichood Hypothesis, whereby they suggest that if the embedded NP refers to the topic of the ongoing discourse, then it tends to be the subject, resulting in an object RC (5). Mak et al. (2008) and Roland et al. (2012) both showed that when provided with appropriate discourse context, that is when the embedded NP of a RC refers to the topic of the ongoing discourse, the difficulty associated with processing object RCs can be reduced in self-paced reading tasks with adults. However, there is a need to partial out whether it is merely the effect of the embedded NP being discourse-old or whether it is a true effect of the embedded NP being the topic of the ongoing discourse. Note that although givenness and topichood are often correlated (i.e., something that is the topic of the ongoing discourse is usually given in the discourse), it is possible that an entity is given, but not the topic of the discourse (e.g., *the sculptor* in (5)).

- (5) **The studio** showed the sculptor a room for rent in the basement. The artist that **the sculptor** admired exhibited portraits at the gallery on Elmwood Avenue. (Roland et al., 2012, p. 493)

Roland et al. (2012) extended these lines of research by manipulating both givenness and topichood of the embedded NP of object RCs in their self-paced reading task. Their results showed that the reading times for object RCs that contain a discourse-old embedded NP referring to the ongoing topic were quicker than those for object RCs containing a discourse-old embedded NP that does not refer to the ongoing topic, as in example (5). These findings suggest that object RCs that meet *both* the givenness and topichood expectations (i.e., the embedded NP is discourse-old and the ongoing topic of the discourse) are easier to process than those that only meet the givenness expectations.

In sum, previous studies with adults have illustrated that the processing of RCs is driven by discourse-based constraints and expectations. Even though there are strong links between formal and discourse-based constraints (e.g., pronouns tend to refer to given referents) and between different types of formal or discourse-based constraints (e.g., the ongoing topic tends to be given), these constraints play individual and distinguishable roles in adult sentence processing. For example, although the embedded NPs in object RCs tend to be both given and expressed by pronouns, when sentences are presented in isolation and the pronoun cannot be linked to a referent mentioned in previous discourse, adults still find it easier to process object RCs when the embedded NP is expressed by a pronoun, as opposed to a lexical NP. Similarly, although the embedded NPs in object RCs tend to be both given and referring to the topic of the ongoing conversation, adults find it easier to process object RCs when *both* of these constraints are met.

Whether children can be expected to follow the same discourse-based cues as adults when they are processing RCs is currently unknown. To our knowledge, previous research has only focused on lexical and semantic features of subject and object RCs produced by children and their caregivers. Whether the RCs in child and child-directed speech contain the same discourse-based features as the RCs found in adult-directed speech (Fox & Thompson, 1990) is an open question, as children (and their caregivers) might not use RCs for the same reasons – or with the same discourse functions – as adults. For example, Diessel and Tomasello (2000, p. 135) have shown that English-speaking children's early RCs are mainly used to characterize NPs (e.g., *You left this toy I'm playing with. Here's a tiger that's gonna scare him*; for similar patterns in German see Brandt et al., 2008). However, whether these NPs are discourse-given or new or whether they refer to the topic of the ongoing conversation is unknown and will be investigated in the current study.

Some studies have shown that children show sensitivity to aspects of discourse context (e.g., givenness and topichood) in their language input early in development, and their comprehension and production of simple linguistic constructions is guided by such information. For example, Matthews et al. (2006) found that 3- to 4-year-olds used more pronouns, compared to lexical NPs, to refer to entities that had already been mentioned in the ongoing conversation. In another study, Song and Fisher (2005) found that 2.5-year-olds tend to disambiguate pronouns as referring to entities in the subject (i.e., topic) position or first-mentioned entities. Yet, disentangling the relations between discourse factors, such as givenness or topichood, and choice of lexical form takes developmental time. For example, 5-year-olds still make errors, using pronominal subjects in simple transitives when the referent *cannot* be readily identified from the prior discourse (Theakston, 2012). Moreover, studies looking at children's processing of RCs traditionally only focus on lexical and semantic factors, omitting the influence of discourse context (e.g., Kidd et al., 2007) or limiting the study of discourse context to the use of different NP types (e.g., lexical NPs vs.

pronouns; Brandt et al., 2009; Haendler et al., 2015).

According to experience-based approaches, it is highly likely that the difficulty associated with processing object RCs in previous experimental studies was also due to the test sentences not meeting the discourse-based expectations a language user has for object RCs, based on their prior experience of the input and their own production, as identified by previous corpus studies (e.g., Fox & Thompson, 1990; Gennari & MacDondald, 2009). In order to more clearly understand the reasons why certain types of object RCs are more difficult to process for children, we need to establish the discourse-based characteristics of RCs in their natural language.

### **Present Study**

The primary aim of the present study was to examine distributional patterns of the lexical, semantic, and discourse-based features of NPs in subject and object RCs in child and child-directed speech. We will also discuss whether our findings align with results from studies focusing on adult-directed speech and written corpus data (Fox & Thompson, 1990; Roland et al., 2012). To our knowledge, our study is the first to consider the role of discourse-level characteristics of NPs in caregivers' and children's production of RCs. We extracted and coded RCs from three densely-collected British English developmental corpora (see Method for details). The lexical and semantic characteristics we coded included the syntactic role of the head NP within the RC (i.e., whether a RC is a subject, object, or oblique RC) and the following attributes of both the head and embedded NPs: animacy (animate, animized inanimate for inanimate objects that can be seen as having agency, such as inanimate objects with facial features, toy humans, toy animals, and toy body parts, or inanimate), type (pronoun, proper noun, or lexical NP), and person and number. For lexical NPs, we further coded whether they were definite, indefinite, bare, or a gerund. For discourse-level characteristics, we coded both the givenness and topichood of the head and embedded NPs. In order to code these discourse-level characteristics, we extracted the 10 utterances preceding each utterance containing a RC (e.g., de Marneffe, et al., 2012; Gundel et al., 1993; Michaelis & Francis, 2007; Prince, 1981). An NP is considered given if it is mentioned or referred to in the preceding 10 utterances, clearly predictable from the context, or if the referent is present in the visual scene; otherwise, it is considered new (see Appendix A for more details). For topichood, we considered whether the NP was the focus of the conversation in the preceding 10 utterances. If there was a switch in topics, the topic closest to the target utterance was considered (see Appendix A for more details). These discourse-level features were included as they have been shown to play a role in the processing of subject and object RCs in previous research with adults. In addition, we coded the syntactic role of the head NP in the main clause and the relativizer. We also coded a number of variables relating to the embedded verb for the purpose of gathering frequency statistics for generating experimental stimuli for future studies.

We first gathered distributional information and frequency statistics on the lexical, semantic, and discourse-level characteristics of the RCs in the corpora. We then examined how different lexical, semantic, and discourse-level variables are related to the head and embedded NPs of subject compared to object RCs. Based on previous corpus analyses, we have specific hypotheses for how these cues might be used together or individually. We predicted the following:

- H1 Lexical form and semantics of embedded NPs in object RCs:** Object RCs are likely to be used with animate and pronominal (especially first- and second-person personal pronouns) embedded NPs (Fox & Thompson, 1990; Haendler et al., 2015; Kidd et al., 2007; Warren & Gibson, 2002), or with definite lexical embedded NPs (Fox & Thompson, 1990; e.g., *Look at all this mess **you**'ve made; But are those the ones that **the man** tried on you?*).
- H2 Discourse-related characteristics of head and embedded NPs in object RCs:** Object RCs are likely to be used with embedded NPs that refer to given referents (Fox & Thompson, 1990; Roland et al., 2012) and the ongoing topic of the discourse context (Roland et al., 2012). In addition, object RCs are likely to be used with head NPs that refer to new referents (Fox & Thompson, 1990; e.g., *Tell me about **the food you** had today at school*).
- H3 Head NPs of object RCs:** Object RCs are likely to be used with head NPs that are inanimate and expressed by definite lexical NPs (Fox & Thompson, 1990; Kidd et al., 2007; e.g., *Look I'm undoing **the knot** I did*).
- H4 Embedded NPs in subject RCs:** Subject RCs are likely to be used with embedded NPs that are indefinite lexical NPs and refer to new referents (Fox & Thompson, 1990; Roland et al., 2012; e.g., *Somewhere that's got **big stairs** in*).
- H5 Head NPs of subject RCs:** Subject RCs are likely to be used with head NPs that are animate and indefinite lexical NPs (Fox & Thompson, 1990; Kidd et al., 2007; e.g., *Little Miss Sunshine discovers **a wicked witch** who is casting evil spells*).

Some of the main lexical and semantic patterns have already been shown in previous analyses of child and child-directed speech. For example, Kidd et al. (2007) have demonstrated that English- and German-speaking children tend to use object RCs with inanimate head NPs and pronominal embedded NPs. One aim of the current study was to replicate these findings. In addition, we conducted a more fine-grained analysis of lexical NPs to distinguish between definite and indefinite forms, which also tend to be associated with discourse-given and discourse-new referents respectively. Most importantly, our corpus analysis also incorporated a separate analysis of the discourse status of NPs in subject and object RCs.

As child and child-directed speech may not necessarily resemble typical adult

language use, we first investigated whether and how the use of subject and object RCs in child and child-directed speech aligns with discourse context as well as lexical and semantic features (e.g., are object RCs associated with discourse-given entities in the embedded-NP slot (i.e., the subject)?) through some preliminary analyses. We used exploratory analytical techniques, including multiple correspondence analysis (MCA) and random forest analysis, to see whether previously identified features of RCs based on typical adult language use are evident in child and child-directed speech. These preliminary analyses were used to *identify* key features of head and embedded NPs in subject and object RCs. We only report a summary of the results from our preliminary analyses using exploratory analytical techniques in the results section; details and full results are available on OSF (see link in the Data Availability section).

We then used chi-squared tests to examine the lexical, semantic, and discourse-level characteristics of NPs in subject and object RCs in child and child-directed speech. We investigated whether any associations between RC type (subject RC vs. object RC) and our coded lexical, semantic, and discourse-level variables, identified via our exploratory analytical techniques, were statistically significant.

We also investigated how well characteristics of subject and object RCs in children's input align with their own RC production. Child and child-directed speech were initially analyzed separately. We expected the RCs in child speech, in general, to show similar characteristics to those identified in child-directed speech. Where we observed differences in the lexical, semantic, or discourse-level factors for child-directed and child speech, we conducted further chi-square tests to test for differences between child-directed and child speech to understand how children's own production of RCs differ from those in their language input.

The present study did not directly test online sentence processing. Rather, it provides a distributional analysis of the lexical, semantic, and discourse characteristics of the head and embedded NPs of subject and object RCs in naturalistic child-directed and child speech. Under experience- and usage-based approaches, such distributional patterns shape the expectations that language users bring to comprehension and production.

## **Method**

### **Data**

Our data came from three high-density British English developmental corpora (Lieven, et al., 2009): Thomas, Gina, and Helen, available on CHILDES (MacWhinney, 2000). The Thomas corpus consists of data from 2;00.12 to 4;11.20 years, with a total of 379 hour-long recordings. The Gina and Helen corpora are smaller. They are part

of the MPI-EVA-Manchester corpus (Lieven et al., 2009). The Gina corpus consists of data from 3;00.01 to 4;07.29 years, with 118 hour-long recordings; whereas the Helen corpus consists of data from 3;00.01 to 5;01.19 years, with 183 hour-long recordings. All three children came from middle-class backgrounds in the North of England, and their primary caregivers were their mothers. These densely-collected data allow us to have a representative picture of the kind of language that children learning British English as their first language are exposed to and produce during everyday play interactions in the home (see Appendix A for details of recording patterns). These corpora were selected because they are among the densest British English developmental corpora available on CHILDES that contain frequent naturalistic recordings across several years of development. Their recording density and conversational richness make them particularly suitable for analyzing relatively low-frequency complex sentences (e.g., RCs) at early stages of language development.

**Table 1. CLAN commands for search string**

| Corpus         | Search string in CLAN                                                                                                                                                                                                                                                                    |
|----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Gina and Helen | combo +t%mor +s"pro:wh who* +pro:wh which* +pro:wh when*<br>+pro:wh where* +pro:wh whose* +pro:wh whom* +rel * +comp that*<br>+pro:dem that* +((pro*+n*)^(pro*+n*)^(v *+cop *+mod *+aux*))<br>+((pro*+n*)^(pro*aux *+n*aux *)) +((pro*+n*)^(det *)^(n*))" *.cha +f +u -w10               |
| Thomas         | combo +t%mor +s"pro:int who* +pro:int which* +pro:int when*<br>+pro:int where* +pro:int whose* +pro:int whom* +pro:rel * +comp that*<br>+pro:dem that* +((pro*+n*)^(pro*+n*)^(v *+cop *+mod *+aux*))<br>+((pro*+n*)^(pro*aux *+n*aux *)) +((pro*+n*)^(det:art *)^(n*))" *.cha +f +u -w10 |

*Note.* The search strings for Gina and Helen and that for Thomas are essentially the same, the differences were due to the original transcription being done in different versions of CLAN.

We first carried out preliminary extraction on a subset of the transcripts to determine our sample plan for RCs in caregiver and child speech and determine the best search string to use across the transcripts. Two individual researchers manually extracted all RCs produced by the children and their caregivers from the two final sessions of each corpus, as these sessions would contain more RCs compared to earlier sessions. We then resolved any inconsistencies in our individual manual extractions – making sure that all identified utterances were actually RCs and all RCs within the selected sessions were identified. We established that the best way to capture all RCs in the corpora was to search for all utterances with (1) a relativizer; (2) the following interrogative/wh- pronouns: who, which, where, whose, and whom; (3) “that” marked as a complementizer or demonstrative pronoun; (4) a noun/pronoun followed by another noun/pronoun followed by a verb/copula/modal verb/auxiliary verb; or (5) a noun/pronoun followed by a determiner/article followed by a noun. Due to minor

differences in the format of the morphological coding tier (%mor), the search strings differed slightly for the corpora – see Table 1 for the search string in CLAN for each corpus. The search strings were then used to extract RCs from the first available session and the middle session of each corpus and the results compared against manually extracted results. This confirmed that the search strings could capture all of the manually identified RCs. We then ran the corresponding search string on each complete corpus.

In order to get a manageable and sufficient sample of RCs for analysis, we used the numbers from our manual extraction, the numbers of hits using the search strings, and the numbers of utterances identified from the whole and a subset of the corpora using the search strings to estimate the number of RCs we could obtain from the corpora. Based on these estimated numbers, we sampled a subset of sessions for caregiver input to get a manageable sample of RCs and used all available sessions for child production. The subset of sessions we focused on for caregiver input were the (child) age ranges where transcripts are available for all three children: the first six weeks starting at age 3;00.01, the first week starting at age 4;00.02, and all weeks between 4;02.29 and 4;07.29.

**Table 2. Details of extracted utterances**

| Corpus | Speaker   | No. of<br>automatically<br>extracted<br>utterances | No. of<br>headed<br>RCs | No. of<br>headless<br>RCs | No. of<br>incomplete<br>RCs |
|--------|-----------|----------------------------------------------------|-------------------------|---------------------------|-----------------------------|
| Gina   | Caregiver | 12843                                              | 227                     | 161                       | 0                           |
|        | Child     | 13633                                              | 79                      | 83                        | 10                          |
| Helen  | Caregiver | 7734                                               | 262                     | 102                       | 9                           |
|        | Child     | 8110                                               | 276                     | 141                       | 35                          |
| Thomas | Caregiver | 10484                                              | 737                     | 205                       | 48                          |
|        | Child     | 19083                                              | 140                     | 114                       | 20                          |

RCs were then manually identified from the automatically extracted utterances. We included 10 lines preceding the target utterance to provide discourse context (Gundel et al., 1993). Relative-clause utterances were excluded if they were sung nursery rhymes, incomplete and syntactically not analyzable, self-repetitions, or did not have a clear head NP (e.g., *this is what we want*)<sup>2</sup>. Overall, we analyzed 1721 utterances (495 from the children, 1226 from the caregivers; see Table 2 for details).

<sup>2</sup> We have omitted headless RCs from our analyses because we were interested in the form and function of head NPs and embedded NPs and how these features are associated with subject versus object RC. Whereas object and oblique RCs can be headless, subject RCs are

## Coding

Each RC utterance was coded for 28 lexical, semantic, and discourse-level variables (see Appendix A for coding scheme). We first marked the head and embedded NP of each RC. Then, we coded the syntactic role of the head NP in the main clause and the relativized NP within the RC (i.e., whether it was a subject, object or oblique RC). We also noted the relativizer of each RC, if present. For the head and embedded NPs, we coded their animacy, type (i.e., pronoun, proper noun, or lexical NP), and person and number. For lexical NPs, we further coded whether they were definite, indefinite, bare, or a gerund. For proper nouns, we further coded whether they were the child's name, a high frequency word (i.e., a noun that accounts for more than 1% of the proper noun tokens in the corpus), or a general proper noun. For discourse-level characteristics, we coded both the givenness<sup>3</sup> and topichood of the head and embedded NPs. Note that even though NPs that refer to the topic of the ongoing conversation are by default also discourse-given, these two discourse-level characteristics were coded and analyzed separately, as they have been shown to play separate roles in language processing (Roland et al., 2012).

We also coded some additional variables that may be of interest to other researchers. Full details of these other coded variables can be found in the coding scheme (Appendix A).

The same coding scheme was used for both the caregiver and child data. The first author and a trained research assistant coded the data. A random sample of the data (~ 20% of the caregiver and child data of each corpus) was second-coded by the first<sup>4</sup> and the third authors for reliability. The agreement across raters was measured using free-marginal multirater kappa ( $\kappa$ ; Randolph, 2005). Unlike other agreement measures like Cohen's kappa (Cohen, 1960) or Fleiss' kappa (Fleiss, 1971), this measure does not assume that the raters know a priori how many cases they should assign to each category of a variable. The average agreement for the lexical and semantic variables was .96 (.97 for caregiver data and .96 for child data; see further

---

always headed. Future research could investigate whether embedded NPs in headless object RCs show the same features as those in headed object RCs.

<sup>3</sup> Initially, we introduced an 'associated' category for givenness to capture noun phrases that are semantically associated with a referent, action, or concept in the preceding 10 utterances (e.g., *ladies* in a previous utterance being referred to as *people* in the relative clause). However, to make interpretation of data more meaningful, we recoded all instances of 'associated' as 'new' and included this coding in our analyses. This change in coding did not affect our results.

<sup>4</sup> The first author only second-coded data that were initially coded by the trained research assistant.

details in Appendix A), and .83 for the discourse-level variables (.86 for caregiver data and .79 for child data; see further details in Appendix A). Disagreements between raters were identified, discussed, and resolved. Whenever a systematic disagreement was identified, the first rater revisited all coded utterances. In the rare cases where the judgements of both raters were deemed appropriate (see Excerpt (6) from Thomas's mother, where the first rater coded the embedded NP *you* as the topic of the ongoing discourse, but the second coder coded it as non-topic as the topic could also be *someone else* or *they*), the judgement of the first rater was used. The kappas reported here are based on the raters' initial coding.

- (6) MOT<sup>5</sup>: As you're growing up and getting too big for some of your toys, it would be nice if you gave them to somebody else, so they could enjoy them.

CHI: Yeah.

MOT: We don't really know many little boys that are younger than you.

## Results

The aim of the present study was to explore which lexical, semantic, and discourse-level features characterize subject and object RCs in both child-directed and child speech. Based on our findings, we should be better able to formulate hypotheses for how children's processing of RCs might be affected by discourse-based cues, in addition to lexical and semantic cues. We first provide descriptive statistics and examples from the corpora analyzed, which will also allow a direct comparison between our findings and patterns previously identified in adult-directed speech (Fox & Thompson, 1990). We then provide a short summary of our preliminary analyses and report the results from chi-squared tests that examined how the distribution of different features of head and embedded NPs differs between subject and object RCs.

### Descriptive Statistics

Across the corpora, there were 422 subject RCs and 586 object RCs in the caregiver data, and 201 subject RCs and 211 object RCs in the child data<sup>6</sup>. More detailed

---

<sup>5</sup> MOT refers to mother and CHI refers to child.

<sup>6</sup> Our sample contained 45 utterances directly read from books. Of these 45 utterances, 28 were subject RCs and 12 were object RCs (with the remaining being oblique RCs). In general, the object RCs were more likely used with an animate head NP and all embedded NPs were animate and expressed by pronouns. The subject RCs were also more likely used with an animate head NP, whereas the embedded NPs were slightly more likely to be inanimate and expressed by lexical NPs. Apart from the head NPs in object RCs being more likely to be animate, other distributional patterns relating to animacy and type of NP were similar to those observed in spontaneous speech.

descriptive statistics and relevant examples can be found in Tables 3 to 6. Please note that the tables only contain the most frequent and prominent patterns. For example, the first row in Table 3 shows how many of the subject and object RCs contained an animate head NP. As displayed in Table 3 both the caregivers and children had a strong tendency to use object RCs with inanimate head NPs (87% and 82% respectively), whereas subject RCs were used with both animate and inanimate head NPs. In the caregiver data, 44% of subject RCs had an animate head, and 50% had an inanimate head. In the child data, 48% of subject RCs had an animate head, and 37% had an inanimate head. The tables do not show any data for oblique RCs<sup>7</sup>. We have also excluded categories with only small numbers of exemplars from the descriptive statistics and overview tables. For example, only a few subject and object RCs were used with head NPs that we coded as animized inanimate (see coding scheme in Appendix A). Also note that a good number of the subject RCs were intransitive, which means that they did not contain an embedded NP. Therefore, the *n*'s for head NPs of subject RCs are higher than the *n*'s for embedded NPs of subject RCs.

The same descriptive statistics as in Tables 3 to 6 but by corpus are provided in Table B.1 in Appendix B. We note that the three corpora showed similar trends in general. Any differences observed pertain to tendencies where only a few utterances were present or tendencies that do not seem to be particularly strong.

---

<sup>7</sup> Although oblique RCs are important for understanding the broader distribution of RC structures, the present study focused specifically on subject and object RCs because the theoretical processing asymmetry literature has primarily contrasted these two structures. In addition, subject and object RCs differ in the syntactic role of the relativized NP itself, whereas oblique RCs introduce additional variability related to prepositional structure and argument realization. We note that most RCs in our data were subject and object RCs, with oblique RCs making up only 17.78% of the RCs produced by caregivers ( $n = 218$ ) and 16.77% of those produced by children ( $n = 83$ ). Nonetheless, oblique RCs constitute an important avenue for future research, particularly given their overlap with object RCs in their earlier linear structure.

**Table 3. Animacy of NPs**

|                       | Child-directed speech                                                                                    |                                                                          | Child speech                                                                       |                                                                     |
|-----------------------|----------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|------------------------------------------------------------------------------------|---------------------------------------------------------------------|
|                       | Subject RC (n = 422)                                                                                     | Object RC (n = 586)                                                      | Subject RC (n = 201)                                                               | Object RC (n = 211)                                                 |
| Head NP animate       | n = 187 (44%)<br><i>Little Miss Sunshine discovers <b>a wicked witch</b> who is casting evil spells.</i> | n = 50 (9%)<br><i>Is that <b>the cow</b> that I can see round there?</i> | n = 96 (48%)<br><i><b>A big lady</b> who's got the girl.</i>                       | n = 18 (9%)<br><i>Look at all <b>the wiggly worms</b> I've got.</i> |
| Head NP inanimate     | n = 213 (50%)<br><i>So which <b>jigsaw pieces</b> can you see that are red?</i>                          | n = 511 (87%)<br><i><b>Something</b> that you really wanna do.</i>       | n = 75 (37%)<br><i>Do you have <b>the thing</b> that goes in here?</i>             | n = 174 (82%)<br><i>And I've got <b>sand</b> that we play with.</i> |
| Head NP ambiguous     | n = 22 (6%)<br><i>Get <b>one</b> that goes bee.</i>                                                      | n = 25 (4%)<br><i>This is another <b>one</b> we do with Julian.</i>      | n = 30 (15%)<br><i>Shall I show you <b>something</b> else that's really funny?</i> | n = 19 (9%)<br><i>Get <b>something</b> I want.</i>                  |
| Embedded NP animate   | n = 73 (17%)<br><i>Was children who interrupt <b>other people</b> when they're talking.</i>              | n = 576 (98%)<br><i>Choose one that <b>we</b> all know.</i>              | n = 37 (18%)<br><i>You're the one that feeds <b>me</b> the food.</i>               | n = 205 (97%)<br><i>Them knickers that <b>I</b> had on today.</i>   |
| Embedded NP inanimate | n = 206 (49%)                                                                                            | n = 10 (2%)                                                              | n = 73 (36%)                                                                       | n = 2 (1%)                                                          |

|                       |                                                                       |                                                |                                                                               |                                                                                  |
|-----------------------|-----------------------------------------------------------------------|------------------------------------------------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
|                       | <i>The ones that don't have <b>edges</b> can go in there.</i>         | <i>It was a suitcase <b>the train</b> hit.</i> | <i>Somewhere that's got <b>big stairs</b> in.</i>                             | <i>I bet you're thinking of trucks that <b>trains</b> carry.</i>                 |
| Embedded NP ambiguous | <i>n = 5 (1%)</i><br><i>And what's that that's next to <b>it</b>?</i> |                                                | <i>n = 8 (4%)</i><br><i>The person that is standing next to <b>them</b>.</i>  | <i>n = 4 (2%)</i><br><i>Have you seen those loading things <b>they</b> have?</i> |
| Embedded NP N/A       | <i>n = 138 (33%)</i><br><i>The wasp that cried out.</i>               |                                                | <i>n = 83 (41%)</i><br><i>Why do they have to wear things that are light?</i> |                                                                                  |

*Note.* The 'ambiguous' category includes 'animized inanimate'. The embedded NPs of intransitive subject RCs were coded as 'N/A'.

**Table 4. Type of NPs**

|                     | Child-directed speech                                                                             |                                                                                          | Child speech                                                                   |                                                                       |
|---------------------|---------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|-----------------------------------------------------------------------|
|                     | Subject RC (n = 422)                                                                              | Object RC (n = 586)                                                                      | Subject RC (n = 201)                                                           | Object RC (n = 211)                                                   |
| Head NP pronoun     | n = 55 (13%)<br><i>Even when I've done <b>something</b> that isn't good.</i>                      | n = 63 (10%)<br><i>Because they might send you <b>something</b> you haven't ordered.</i> | n = 32 (16%)<br><i>It's <b>me</b> who's going to get sticky fingers.</i>       | n = 46 (22%)<br><i><b>Him</b> over there that I wanna know.</i>       |
| Head NP proper noun | n = 22 (5%)<br><i><b>Queen Mary</b> who wasn't very well.</i>                                     | n = 5 (1%)<br><i>This is <b>Freddy</b> who Thomas takes to bed.</i>                      | n = 6 (3%)<br><i>You know (<b>friend's name</b>) who's very old.</i>           |                                                                       |
| Head NP lexical NP  | n = 343 (81%)<br><i>Shall I be <b>the vet</b> who looks at the animals?</i>                       | n = 514 (88%)<br><i>Tell me about <b>the food</b> that you had today at school.</i>      | n = 161 (80%)<br><i>We can't find <b>the tops</b> that go over that.</i>       | n = 165 (78%)<br><i>Look I'm undoing <b>the knot</b> I did.</i>       |
| Head NP ambiguous   | n = 2 (1%)<br><i>It's <b>Little Bear and the bunny rabbit</b> that have gone up to the attic.</i> | n = 4 (1%)<br><i><b>The t-shirts and things</b> that they were selling.</i>              | n = 2 (1%)<br><i>And that's <b>the spoon and the dish</b> who ran.</i>         |                                                                       |
| Embedded NP pronoun | n = 56 (13%)<br><i>That's the lady that helps <b>him</b>.</i>                                     | n = 509 (87%)<br><i>Look at all this mess <b>you</b>'ve made.</i>                        | n = 25 (13%)<br><i>The library teacher who takes <b>us</b> to the library.</i> | n = 200 (95%)<br><i>Another batterys <b>I</b> can't get out here.</i> |

|                            |                                                                       |                                                                                       |                                                                              |                                                                                       |
|----------------------------|-----------------------------------------------------------------------|---------------------------------------------------------------------------------------|------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| Embedded NP<br>proper noun | <i>n</i> = 13 (3%)<br><i>The people who look after <b>Ginger</b>.</i> | <i>n</i> = 59 (10%)<br><i>Is that the lorry that <b>Granddad</b> bought you?</i>      | <i>n</i> = 7 (3%)<br><i>One Grandad that live with <b>Auntie_(name)</b>.</i> | <i>n</i> = 5 (2%)<br><i>Loads of Smarties that <b>Father Christmas</b> bought me.</i> |
| Embedded NP<br>lexical NP  | <i>n</i> = 215 (51%)<br><i>Is it boys that rip <b>things</b>?</i>     | <i>n</i> = 12 (2%)<br><i>But are those the ones that <b>the man</b> tried on you?</i> | <i>n</i> = 84 (42%)<br><i>The man that that sells um <b>the cows</b>.</i>    | <i>n</i> = 6 (3%)<br><i>It's the thing what <b>men</b> do on the river.</i>           |
| Embedded NP<br>ambiguous   |                                                                       | <i>n</i> = 6 (1%)<br><i>The food that <b>me and Helen</b> were gonna have.</i>        | <i>n</i> = 2 (1%)<br><i>Got a flower one that says <b>you and Daddy</b>.</i> |                                                                                       |
| Embedded NP<br>N/A         | <i>n</i> = 138 (33%)<br><i>The wasp that cried out.</i>               |                                                                                       | <i>n</i> = 83 (41%)<br><i>I'll get the ones that are the same.</i>           |                                                                                       |

**Table 5. Givenness of NPs**

|                     | Child-directed speech        |                             | Child speech                 |                             |
|---------------------|------------------------------|-----------------------------|------------------------------|-----------------------------|
|                     | Subject RC ( <i>n</i> = 422) | Object RC ( <i>n</i> = 586) | Subject RC ( <i>n</i> = 201) | Object RC ( <i>n</i> = 211) |
| Head NP given       | <i>n</i> = 223 (53%)         | <i>n</i> = 321 (55%)        | <i>n</i> = 116 (58%)         | <i>n</i> = 114 (54%)        |
| Head NP new         | <i>n</i> = 145 (34%)         | <i>n</i> = 183 (31%)        | <i>n</i> = 55 (27%)          | <i>n</i> = 58 (27%)         |
| Head NP unclear     | <i>n</i> = 54 (13%)          | <i>n</i> = 82 (14%)         | <i>n</i> = 30 (15%)          | <i>n</i> = 39 (19%)         |
|                     |                              |                             |                              |                             |
| Embedded NP given   | <i>n</i> = 127 (30%)         | <i>n</i> = 530 (90%)        | <i>n</i> = 50 (25%)          | <i>n</i> = 196 (93%)        |
| Embedded NP new     | <i>n</i> = 112 (27%)         | <i>n</i> = 38 (7%)          | <i>n</i> = 46 (23%)          | <i>n</i> = 11 (5%)          |
| Embedded NP unclear | <i>n</i> = 45 (10%)          | <i>n</i> = 18 (3%)          | <i>n</i> = 22 (11%)          | <i>n</i> = 4 (2%)           |
| Embedded NP N/A     | <i>n</i> = 138 (33%)         |                             | <i>n</i> = 83 (41%)          |                             |

**Table 6. Topichood of NPs**

|                       | Child-directed speech        |                             | Child speech                 |                             |
|-----------------------|------------------------------|-----------------------------|------------------------------|-----------------------------|
|                       | Subject RC ( <i>n</i> = 422) | Object RC ( <i>n</i> = 586) | Subject RC ( <i>n</i> = 201) | Object RC ( <i>n</i> = 211) |
| Head NP topic         | <i>n</i> = 152 (36%)         | <i>n</i> = 224 (38%)        | <i>n</i> = 60 (30%)          | <i>n</i> = 70 (33%)         |
| Head NP non-topic     | <i>n</i> = 236 (56%)         | <i>n</i> = 298 (51%)        | <i>n</i> = 122 (61%)         | <i>n</i> = 119 (56%)        |
| Head NP unclear       | <i>n</i> = 34 (8%)           | <i>n</i> = 64 (11%)         | <i>n</i> = 19 (9%)           | <i>n</i> = 22 (11%)         |
|                       |                              |                             |                              |                             |
| Embedded NP topic     | <i>n</i> = 36 (9%)           | <i>n</i> = 20 (3%)          | <i>n</i> = 11 (6%)           | <i>n</i> = 11 (5%)          |
| Embedded NP non-topic | <i>n</i> = 242 (57%)         | <i>n</i> = 532 (91%)        | <i>n</i> = 101 (50%)         | <i>n</i> = 185 (88%)        |
| Embedded NP unclear   | <i>n</i> = 6 (1%)            | <i>n</i> = 34 (6%)          | <i>n</i> = 6 (3%)            | <i>n</i> = 15 (7%)          |
| Embedded NP N/A       | <i>n</i> = 138 (33%)         |                             | <i>n</i> = 83 (41%)          |                             |

***Examples from Thomas' Mother Illustrating Givenness of NPs***(7) Subject RC – Head NP given

MOT: Do you know the other thing I'm going to pop on the grill pan, Thomas?

CHI: Yes.

MOT: This banana that has been lying here all day.

(8) Object RC – Head NP given

MOT: And what will be inside your blue parcel?

CHI: I think a big big one fire engine.

MOT: A big big big fire engine.

MOT: You mean larger than the fire engines you've got now.

(9) Subject RC – Embedded NP given

MOT: I like easter egg chocolate.

CHI: I need it.

CHI: Some more please.

MOT: You...

CHI: More.

MOT: You still have chocolate, Thomas.

MOT: Eat it up quickly.

MOT: There is other kinds of chocolate.

MOT: Right.

MOT: Shall we have one of these that's inside the egg?

(10) Object RC – Embedded NP given

MOT: Well those are flowers that I carried when Aunty Trisha got married.

***Examples from Thomas' Mother Illustrating Non-Topic NPs:***(11) Subject RC – Head NP non-topic

MOT: Or who else likes **bread**, Thomas?

CHI: Don't know.

CHI: Crocodiles?

MOT: Crocodiles?

MOT: Mhm, don't get a lot of crocodiles round here.

MOT: The ducks?

CHI: Yeah.

MOT: Last week when we went to Fielden Park, dou you remember all the **toast** that was left?

- (12) Object RC – Head NP non-topic  
 MOT: So do you think perhaps Grandad um would read about **Teddy Bright Eyes and Tubby Toes** [/] **Teddy Tubby Toes**?  
 MOT: Do you think he would?  
 CHI: Don't know.  
 CHI: I can't see.  
 MOT: Ah now these are **the bears** that Mummy recognizes.
- (13) Subject RC – Embedded NP non-topic  
 MOT: That's the newspaper fallen out of **your TNT lorry** on the table.  
 MOT: And that's how you like to eat pears but...  
 CHI: So what's that?  
 MOT: Let's have a look .  
 MOT: I think it's just a little hook that hooks **the lorry** onto the cab.
- (14) Object RC – Embedded NP non-topic  
 MOT: We've got Rick Krispies or cornflakes.  
 CHI: Pah.  
 MOT: Or cereal with nuts in that **Mummy** likes but I don't think you will.

Since person and number as well as definiteness are only relevant for certain types of NPs (e.g., definiteness is only relevant for lexical NPs), we do not present tables of descriptives with examples. Instead, we provide a summary of the key features here. Lexical NPs and pronouns in the subject RCs produced by the caregivers ( $n = 422$ ) were predominantly third-person singular ( $n = 308$  for head NPs and  $n = 212$  for embedded NPs) or third-person plural ( $n = 110$  for head NPs and  $n = 48$  in embedded NPs). For object RCs ( $n = 586$ ), a similar pattern was observed for head NPs ( $n = 409$  for third-person singular and  $n = 176$  for third-person plural). In contrast, the embedded NPs in the object RCs were more varied, with pronouns frequently being of second-person singular ( $n = 238$ ), followed by first-person plural ( $n = 138$ ), and first-person singular ( $n = 79$ ). There were also considerable instances of embedded lexical NPs and pronouns in the object RCs being third-person singular ( $n = 104$ ).

Similarly, lexical NPs and pronouns in the subject RCs produced by the children ( $n = 201$ ) were predominantly third-person singular ( $n = 141$  for head NP and  $n = 80$  for embedded NP) and third-person plural ( $n = 58$  for head NP and  $n = 30$  for embedded NP). For object RCs ( $n = 211$ ), a similar pattern was, again, observed for head NPs ( $n = 150$  for third-person singular and  $n = 60$  for third-person plural). However, the embedded NPs in the object RCs were dominated by first-person singular ( $n = 98$ ) and second-person singular ( $n = 51$ ) pronouns, with smaller proportions of first-person plural pronouns ( $n = 29$ ) and third-person singular lexical NPs and pronouns ( $n = 19$ ).

In the caregiver data, both subject RCs ( $n = 343$ ) and object RCs ( $n = 514$ ) were frequently used with lexical head NPs. Both subject and object RCs tended to be used with definite lexical head NPs ( $n = 217$  for subject RCs and  $n = 388$  for object RCs) rather than indefinite lexical head NPs ( $n = 126$  for subject RCs and  $n = 123$  for object RCs). In contrast, only subject RCs were frequently used with lexical NPs as embedded NPs ( $n = 215$ ). Of these, 120 were definite, and 83 were indefinite, therefore also showing a (slight) preference for definite lexical NPs.

Similarly, in the child data, both subject RCs ( $n = 161$ ) and object RCs ( $n = 165$ ) were frequently used with lexical head NPs. Again, both subject and object RCs tended to be used with definite lexical head NPs ( $n = 110$  for subject RCs and  $n = 124$  for object RCs) rather than indefinite lexical head NPs ( $n = 51$  for subject RCs and  $n = 39$  for object RCs). As with the caregiver data, only subject RCs were frequently used with lexical NPs as embedded NPs ( $n = 84$ ). Of these, 48 were definite, and 31 were indefinite, therefore also showing a (slight) preference for definite lexical NPs.

### **Preliminary Analyses**

Here, we provide a short summary of our preliminary analyses using exploratory analytical techniques. Full details of these analyses can be found on OSF (see link in the Data Availability section). The main purpose of these preliminary analyses was to explore whether RCs produced by children and their caregivers contain the same features as the RCs found in adult-directed speech and written corpora (Fox & Thompson, 1990; Roland et al., 2012). Some of these features, such as definiteness of head and embedded NPs, have not been explored in any previous studies investigating children's and caregivers' production of RCs. In addition, our descriptive statistics summarized above already indicate that the RCs produced by children and caregivers do not always show the same discourse-based features as the ones identified in adult-directed speech and written corpora (see Tables 5 and 6).

Our random forest analyses reveal that the following attributes of embedded NPs differ between subject and object RCs: animacy, type, person and number, type of lexical NP, givenness, and topichood. For example, compared to subject RCs, object RCs are more likely to contain an animate embedded NP (see Table 3). In addition, animacy and person and number of the head NP differ between subject and object RCs. For example, compared to subject RCs, object RCs are more likely to occur with an inanimate head NP (see Table 3). These variables consistently rank highly in our analyses across all corpora (for both child and caregiver data). Our MCAs also suggest that type, definiteness (i.e., type of lexical NP), givenness, and topichood of the head NP may differ between subject and object RCs. We conducted a chi-squared test of association on the raw frequencies reported in the Descriptive Statistics section for each of these variables, separately for caregiver and child speech, to look at distributional differences between subject and object RCs. Where the associations

observed between subject and object RCs differed in caregiver and child speech, we conducted additional chi-squared tests to directly compare the distributions between caregiver and child speech.

### Chi-Squared Tests

A summary of all the chi-squared tests of association is presented in Table 7. For all chi-squared tests, we excluded categories where the expected frequencies for both subject and object RCs were below 5 and where categories (e.g., other) were not meaningful for our hypotheses and interpretation. For significant tests, standardized residuals (SRs) exceeding  $\pm 1.96$  were interpreted as indicating significant deviations from expected frequencies at the .05 level.

In line with H1, both the caregivers and children showed a strong tendency to produce object RCs with animate (SR = 22.88 for caregivers, SR = 13.04 for children; see Table 3) and pronominal embedded NPs (SR = 22.57 for caregivers, SR = 13.86 for children; see Table 4) more often than expected and relative to subject RCs. In addition, these pronominal embedded NPs tended to be first-person singular (SR = 5.13 for caregivers, SR = 7.95 for children), first-person plural (SR = 8.48 for caregivers, SR = 3.92 for children), or second-person singular pronouns (SR = 10.94 for caregivers, SR = 5.11 for children).

In line with H2, both the caregivers (SR = 12.97) and children (SR = 9.06) showed a strong tendency to produce object RCs where the embedded NP was given (see Table 5) more often than expected and relative to subject RCs. However, in contrast to H2, despite being discourse-old, the embedded NPs of object RCs were unlikely to refer to the topic of the ongoing conversation for the caregivers (SR = 5.06; see Table 6). A similar but not statistically significant trend was observed for the children (SR = 1.39; see Table 6). Another observation (although not statistically significant) that contradicted H2 was that the caregivers were more likely to produce object RCs with discourse-given (SR = 0.92) rather than discourse-new head NPs, whereas no clear preference was observed for the children (SR = -0.16; see Table 5).

In line with H3, both the caregivers (SR = 13.41) and children (SR = 9.58) showed a strong tendency to produce object RCs with inanimate head NPs (see Table 3) more often than expected and relative to subject RCs. Further, also consistent with H3, the caregivers tended to produce object RCs with definite lexical head NPs (SR = 3.99) more often than expected and relative to subject RCs. The same trend, although not statistically significant, was also observed in the child data (SR = 1.64).

**Table 7. Summary of chi-squared tests of association**

| Variable             | Speaker   | Head NP                                 | Embedded NP                                  |
|----------------------|-----------|-----------------------------------------|----------------------------------------------|
| NP Type              | Caregiver | $\chi^2(1) = 1.86, p = .172, V = .04^*$ | $\chi^2(1) = 509.47, p < .001, V = .81^{**}$ |
|                      | Child     | $\chi^2(1) = 1.76, p = .184, V = .07$   | $\chi^2(1) = 192.05, p < .001, V = .78^{**}$ |
| Animacy              | Caregiver | $\chi^2(1) = 179.93, p < .001, V = .43$ | $\chi^2(1) = 523.38, p < .001, V = .78$      |
|                      | Child     | $\chi^2(1) = 91.82, p < .001, V = .50$  | $\chi^2(1) = 170.09, p < .001, V = .73$      |
| Person and Number*** | Caregiver | $\chi^2(1) = 1.70, p = .192, V = .04$   | $\chi^2(4) = 376.38, p < .001, V = .66$      |
|                      | Child     | $\chi^2(1) = 0.02, p = .898, V = .01$   | $\chi^2(4) = 184.90, p < .001, V = .75$      |
| Definiteness****     | Caregiver | $\chi^2(1) = 15.94, p < .001, V = .14$  | $\chi^2(2) = 2.83, p = .244, V = .11$        |
|                      | Child     | $\chi^2(1) = 2.69, p = .101, V = .09$   | $\chi^2(2) = 1.49, p = .474, V = .13$        |
| Givenness            | Caregiver | $\chi^2(1) = 0.84, p = .360, V = .03$   | $\chi^2(1) = 168.17, p < .001, V = .46$      |
|                      | Child     | $\chi^2(1) = 0.02, p = .876, V = .01$   | $\chi^2(1) = 82.17, p < .001, V = .53$       |
| Topichood            | Caregiver | $\chi^2(1) = 1.28, p = .258, V = .04$   | $\chi^2(1) = 25.56, p < .001, V = .18$       |
|                      | Child     | $\chi^2(1) = 0.68, p = .411, V = .04$   | $\chi^2(1) = 1.94, p = .164, V = .08$        |

\* Test excluded proper nouns for comparison with the child data of which none of the RCs contained a proper noun as head NP.

\*\* Test excluded proper nouns due to the very small number of RCs with a proper noun as an embedded NP (relative to lexical NP and pronoun). This is to avoid making strong claims based on small numbers of utterances where small differences may be significant.

\*\*\* For head NP, the categories included were third-person singular and plural only. For embedded NP, the categories included were first- and third-person singular and plural and second-person singular. The exclusion of other categories was due to low expected frequencies ( $< 5$ ).

\*\*\*\* For head NP, the categories included were definite and indefinite lexical NPs. For embedded NP, bare lexical NPs were included. The exclusion of other categories was due to low expected frequencies ( $< 5$ ).

Note. Reported p-values have not been adjusted for multiple comparisons, but we note that all significant p-values remain significant after Bonferroni correction (all  $ps < .024$ ).

Inconsistent with H4, the association between definiteness of embedded NPs and RC type was not significant. However, consistent with H4, both the caregivers (SR = -12.97) and children (SR = -9.06) tended to use subject RCs with an embedded NP that was discourse-new (see Table 5) more often than expected and relative to object RCs.

Finally, consistent with H5, both the caregivers (SR = -13.41) and children (SR = -9.58) showed a tendency to use subject RCs with an animate head NP (see Table 3) more often than expected and relative to object RCs. In addition, for caregivers, subject RCs were more likely to be used with indefinite lexical head NPs (SR = 3.99), whereas a similar but not statistically significant trend was observed for children (SR = 1.64).

In addition, we also found the following trends that did not speak directly to our hypotheses and were explored based on the results of our preliminary analyses. These trends aligned with those displayed in the descriptive statistics (Tables 3 to 6). The caregivers produced object RCs with the following features less often than expected and relative to subject RCs: an inanimate embedded NP (SR = -22.88; see Table 3), a lexical embedded NP (SR = -22.57; see Table 4), a third-person singular embedded NP (SR = -16.29), a third-person plural embedded NP (SR = -6.52), or a topical embedded NP (SR = -5.06; see Table 6). The children produced object RCs with the following features less than expected: an inanimate embedded NP (SR = -13.04, see Table 3), a lexical embedded NP (SR = -13.86; see Table 4), a third-person singular embedded NP (SR = -11.08), or a third-person plural embedded NP (SR = -4.77).

Of note, examining all chi-squared tests of association addressing our hypotheses and those that followed our preliminary analyses, there were distributional differences between caregiver and child speech for definiteness of head NP and topichood of embedded NP (Table 7). More specifically, the distribution of definite and indefinite head NP and topic and non-topic embedded NP across subject and object RCs differed between caregiver and child speech. We further examined these distributional differences using chi-squared tests of association. We found that the caregivers produced object RC with a definite lexical head NP ( $\chi^2(1) = 8.81, p = .003, V = .10, SR = 2.97$ ) more often than expected relative to subject RC when compared to the children. The distributional differences of indefinite head NP and topic and non-topic embedded NP across subject and object RCs between caregiver and child speech were not statistically different.

## Discussion

The aim of the study was to examine which lexical, semantic, and discourse-level features are present in subject and object RCs in both child-directed and child speech. Based on our findings, we will be better able to formulate hypotheses for how children's processing of subject and object RCs might be affected by discourse-based cues, in addition to lexical and semantic cues, and whether these cues are similar to

the ones observed in the RCs produced in adult-directed speech or found in written corpora (e.g., Fox & Thompson, 1990; Gennari & MacDonald, 2009; Roland, 2012). We had five hypotheses.

H1: We hypothesized that object RCs are most likely to be used with animate and pronominal (especially first- and second-person personal pronouns) embedded NPs or with definite lexical embedded NPs. The descriptive statistics and chi-squared tests are in support of this hypothesis for both child-directed and child speech.

H2: We hypothesized that object RCs are most likely to be used with given and topical embedded NPs. In addition, object RCs are mostly used with new head NPs. We found partial support for this hypothesis. Our descriptive statistics and chi-squared tests showed that the embedded NPs in object RCs likely refer to given referents in child-directed and child speech. Contrary to H2, we found that object RCs were most likely to be used with head NPs that refer to given referents in child-directed speech, whereas a clear preference was not observed in child speech. Furthermore, our chi-squared tests on child-directed speech and the descriptive statistics for both child and child-directed speech showed that object RCs are rarely used with embedded NPs that refer to the ongoing topic of the conversation.

H3: We hypothesized that object RCs are mostly used with inanimate head and definite lexical head NPs. The results of our descriptive statistics and chi-squared tests are in line with this prediction, with the exception that the trend for definite lexical head NPs in object RCs did not reach statistical significance for child speech. Interestingly, the distributional difference between child-directed and child speech was statistically significant for definiteness of lexical head NPs.

H4: We hypothesized that subject RCs are mostly used with indefinite lexical embedded NPs referring to new referents. We found support for this only in our descriptive statistics for both the caregiver and child data. Even though only 24% of caregivers' and 22% of children's subject RCs contained a discourse-new embedded NP, these proportions were much higher than for object RCs, where only 6% and 4% of caregivers' and children's object RCs contained a discourse-new embedded NP. However, these trends did not reach statistical significance in our chi-squared tests.

H5: We hypothesized that subject RCs are mostly used with animate and indefinite lexical head NPs. The results of our descriptive statistics and chi-squared tests provided support for this. Compared to object RCs, subject RCs are much more likely to contain an animate head NP. We also found that subject RCs are more likely to be used with indefinite lexical head NPs, although the trend did not reach statistical significance in child speech.

Robust findings relating to lexical form (e.g., lexical vs. pronominal NP, and definite vs. indefinite lexical NP) and semantics (i.e., animacy) of the head and embedded NPs of subject and object RCs that have been consistently reported in previous experimental and corpus studies (Fox & Thompson, 1990; Haendler et al., 2015; Kidd et al., 2007; Warren & Gibson, 2002) were replicated in our analyses, although the trends for definiteness did not always reach statistical significance. This suggests that the RCs produced by caregivers and children resemble RCs typically produced by adults and in adult-directed speech. More importantly, patterns that have only been observed in adult data in previous studies (e.g., definiteness of embedded NPs in object RCs; Fox & Thompson, 1990; Warren & Gibson, 2002) also seem to apply, at least descriptively, to child and child-directed speech. Taken together, our results and results from previous studies suggest that by 3 years of age, children show similar associations between RC type (subject vs. object RC) and lexical and semantic properties of the head and embedded NPs as adults. However, it is important to note that this lack of developmental differences is likely to reflect an earlier process of learning from the input, such that children at this stage of development (3-5 years old) have already acquired the formal expectations for head and embedded NPs in RCs, suggesting an early sensitivity to the lexical and semantic properties. Interestingly, there were slight differences between child-directed and child speech in terms of definiteness of lexical head NPs in object RCs. Although showing the same trend descriptively, the tendency for caregivers, but not children, to use definite lexical head NPs in object RCs compared to subject RCs was statistically significant. This suggests that children may still be developing an expectation for an association between definiteness of head NPs and types of RCs. We note that this could be due to definiteness being closely related to discourse context, as definite NPs often refer to entities that are given or the topic of the ongoing discourse. As we discuss below, discourse expectations seem to be more complex and could, therefore, be acquired later. As we only looked at how similar and/or different RCs are in child vs. child-directed speech and did not predict child production from their language input, it would be interesting to investigate how – and at what point – children’s production of different types of RCs aligns with their earlier input. It was not possible to investigate this in the present study due to the small number of utterances (relative to the number of coded variables).

In contrast, some of our findings in relation to discourse-level characteristics of head and embedded NPs in subject and object RCs are inconsistent with findings from previous literature. Based on adult spoken and written language, Fox and Thompson (1990) and Roland et al. (2012) suggested a trend for the embedded NP of an object RC to be given and the topic of the ongoing discourse and the head NP of an object RC and the embedded NP of a subject RC to be new. We observed similar patterns to those in previous literature for givenness of embedded NPs in object RCs, but not for topichood of embedded NPs in object RCs or givenness of head NPs in subject RCs, where we either observed the reverse pattern or no clear preferences.

Inconsistent with previous findings, our chi-squared tests on the caregiver data showed that the embedded NP in object RCs tends *not* to be the topic of the previous discourse context. The same pattern was evident in the descriptive statistics for both caregivers and children. We attribute this inconsistency to differences in the conceptualization and operationalization of topichood in previous experimental studies/our corpus study and how topichood actually works in naturalistic discourse. To the best of our knowledge, our study is the first analysis of RCs in spoken discourse that has coded topichood and tried to distinguish between givenness and topichood. Although givenness and topichood are related (e.g., Arnold et al., 2000; Clark & Clark, 1977; Givón, 1983), Roland et al. (2012) have shown that they contribute individually to the processing of object RCs in the written modality. In the present study, we considered the topic immediately before the target utterance containing the RC as the topic of the ongoing discourse context. This operationalization provided us with a systematic way to code topichood, but it may not truly reflect the broader ongoing topic of the discourse. For example, if the ongoing discourse was about a dog, but immediately before the RC target utterance, there was a change of topic to talk about the child, the child would be considered the topic, and the dog non-topic. Without tracking and following the entire corpus, it is difficult to track the broader topic of the ongoing discourse, which may be more salient in the discourse and therefore show stronger associations with type of RC. The only previous experimental work on the influence of topichood on the processing of RCs is by Roland et al. (2012), where they carefully manipulated a short discourse of one sentence, in which the topic was always unambiguous. Yet, in naturalistic conversations, it is highly unlikely to achieve such controlled discourse contexts. In fact, Givón (1983) stated that topichood is partly measured by recency of mention, suggesting that topichood could be a graded, rather than a binary, variable, such that an entity could be the topic immediately before a target utterance, the topic of the broader ongoing discourse, or non-topic in previous discourse. Although such coding of topichood may be difficult, if not impossible, for corpus studies, future experimental work could investigate whether local and global topichood of embedded NPs influence the processing of RCs differently.

Another unexpected finding was that both children and caregivers were more likely to use object RCs with given rather than new head NPs, whereas Fox & Thompson (1990) suggested that, in adult-directed speech, object RCs are used to anchor *new* head NPs into the discourse context. Our finding indicates that object RCs might play different roles in child and child-directed speech compared to adult-directed speech. Children and caregivers might generally be more likely to talk about given referents (i.e., referents that have either been mentioned before and/or are visible in the visual context; Vihman, 2015; Soderstrom, 2007). This, in turn, could mean that like subject RCs, object RCs are also used to characterize and provide additional information about the head NP rather than anchoring it to a given referent (i.e., the embedded NP

in object RCs). Some evidence for this can be seen in the examples provided in Tables 3 and 4, where a few object RCs are even used with pronominal head NPs (e.g., *Choose **one** that we all know; **Something** that you really wanna do*). And similar findings have been discussed in earlier corpus studies investigating the form and function of RCs produced by English- and German-speaking children. These corpus analyses have found that children's early RCs are mainly used to provide additional information about an NP that is introduced in a copular / presentational main clause, such as *here's a tiger that's gonna scare him* (Brandt et al., 2008; Diessel & Tomasello, 2000).

In addition, our finding that both the head NP and the embedded NP in object RCs in caregiver speech tend to be given could also be explained by MacWhinney and Pléh's (1988) suggestion that object RCs are easier to process when they involve fewer perspective shifts. If an object RC was used with only lexical and new NPs (e.g., *The horse that the cow chased jumped over the pig*), this would involve the activation of several referents and perspective shifts between them. Based on the assumption that we prefer to take the perspective of agents (or subjects) when we process sentences, we would have to shift our perspective three times (from horse to cow and back to horse). In contrast, when object RCs contain pronominal and given NPs (e.g., *Choose **one** that **we** all know*), we can assume that all referents (*one* and *we*) are already activated and do not need to come and go in and out of our focus of attention.

Apart from some discourse-based features, our analyses of child and child-directed speech have shown that children and caregivers produce RCs that contain similar lexical and semantic cues that have been found in adult-directed speech. In addition, the features of children's RCs were largely in line with the features of their caregivers' RCs. For example, both children and caregivers were found to produce object RCs with pronominal and given embedded NPs. The findings presented in the current paper allow us to formulate hypotheses for how discourse-based constraints might affect children's processing of RCs. If, for example, future experimental studies find that, like adults (Roland et al., 2012), children are better at processing object RCs that contain discourse-given embedded NPs, this could be interpreted as supporting evidence for experience- and usage-based theories of language acquisition and processing. However, if these discourse-given NPs are expressed by pronouns, as they usually are, this finding would also be compatible with alternative theories. For example, Rizzi's (1990) theory of Relativized Minimality suggests that if the head NP and the gap is intervened by another NP, as is the case for object RCs, the dependency between the head NP and the gap is more difficult to establish (Friedman et al., 2009), and that this is especially true if the intervening NP is of the same form (e.g., the head and embedded NPs are both lexical NPs, or both pronouns; see also Friedmann et al., 2009; Gordon et al., 2001). The problem is that lexical form is related to discourse function. Therefore, in order to demonstrate that children's RC processing is driven by their sensitivity to discourse-level features derived from their language experience, one would have to create test sentences that tease apart form and

function. For example, if children were also better at processing object RCs with lexical head NPs and lexical and discourse-given embedded NPs (compared to lexical and discourse-new embedded NPs), this would provide better evidence for the usage- and experience-based approach, because it could not be caused by just formal differences between the head and embedded NP.

Although the present study did not directly examine online processing, the distribution patterns observed provide important hypotheses regarding the kinds of object RCs children are likely to process successfully. Both caregivers and children produced object RCs that show similar lexical, semantic, and discourse features, particularly with respect to embedded NPs. Our analyzed object RCs mostly contained animate, discourse-given, and pronominal embedded NPs, often realized as first- or second-person pronouns. Such distributions may help reduce processing demands by increasing predictability, which would further support experience-based theories of language processing. This hypothesis will need to be tested in future studies manipulating not only the semantic and lexical, but also the discourse-based features of the NPs used in RCs. Importantly, our findings do not contradict structural or memory-based accounts of RC processing. Object RCs still involve longer dependencies and intervention configurations than subject RCs. However, the distribution patterns in the corpora suggest that children typically encounter object RCs with specific configurations, with highly predictable lexical, semantic, and discourse cues. This supports the view that structural complexity and distributional patterns jointly shape processing difficulty (Futrell et al., 2020), such that structural locality constraints interact with probabilistic expectations derived from language experience.

Another potentially important source of children's experience with RCs is shared book reading. Previous work (e.g., Hsiao et al., 2022) has shown that children's books often contain a larger proportion of RCs than child-directed speech. Exposure to child-directed texts may therefore increase children's experience with structurally complex RCs and potentially provide more cues to expectations for how lexical, semantic, and discourse features of head and embedded NPs align in different types of RCs. Future work could compare discourse patterns across child-directed speech and child-directed texts to see if the distributional patterns are similar across both types of input and whether higher exposure to child-directed texts supports children's integration of lexical, semantic, and discourse cues for the processing of RCs.

In sum, both children's and adults' RCs display discourse-based features (i.e., givenness and topichood), and some of these discourse-based features are different from the ones found in RCs in adult-directed speech and written corpora. Nevertheless, children produce subject and object RCs with discourse characteristics that resemble those in their input. Further, our results also suggest that children develop an early adult-like sensitivity to the lexical and semantic features of different

types of RCs. Our findings suggest that experimental studies examining the processing of RCs should also take discourse-level factors into account, and that this could help us to distinguish between experience-based and structural complexity and processing difficulty accounts.

### References

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2), 239-273. <https://doi.org/10.1017/S030500091400049X>

Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1), 28-55. <https://doi.org/10.1353/lan.2000.0045>

Bever, T. G. (1970). The cognitive basis for linguistic structures. In R. Hayes (Ed.), *Cognition and language development* (pp. 279-362). Wiley & Sons.

Brandt, S., Diessel, H., & Tomasello, M. (2008). The acquisition of German relative clauses: A case study. *Journal of Child Language*, 35(2), 325-348. <https://doi.org/10.1017/S0305000907008379>

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*, 20(3), 539-570. <https://doi.org/10.1515/COGL.2009.024>

Brandt, S., Lieven, E., & Tomasello, M. (2016). German children's use of word order and case marking to interpret simple and complex sentences: Testing differences between constructions and lexical items. *Language Learning and Development*, 12(2), 156-182. <https://doi.org/10.1080/15475441.2015.1052448>

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. Harcourt Brace Jovanovich.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46. <https://doi.org/10.1177/001316446002000104>

de Marneffe, M. C., Grimm, S., Arnon, I., Kirby, S., & Bresnan, J. (2012). A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes*, 27(1), 25-61. <https://doi.org/10.1080/01690965.2010.542651>

- de Villiers, J. G., Tager Flusberg, H. B., Hakuta, K., & Cohen, M. (1979). Children's comprehension of relative clauses. *Journal of Psycholinguistic Research*, 8, 499-518. <https://doi.org/10.1007/bf01067332>
- Diessel, H., & Tomasello, M. (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics*, 11(1-2), 131-151. <https://doi.org/10.1515/cogl.2001.006>
- Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, 81(4), 882-906. <https://www.jstor.org/stable/4490021>
- Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25(2), 108-127. <https://doi.org/10.1016/j.newideapsych.2007.02.002>
- Du Bois, J. W. (1987). The discourse basis of ergativity. *Language*, 63(4), 805-855. <https://doi.org/10.2307/415719>
- Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2), 143-188. <https://doi.org/10.1017/S0272263102002024>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382. <https://doi.org/10.1037/h0031619>
- Fox, B. A., & Thompson, S. A. (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language*, 66(2), 297-316. <https://doi.org/10.2307/414888>
- Frazier, L., & Clifton Jr, C. (1989). Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes*, 4(2), 93-126. <https://doi.org/10.1080/01690968908406359>
- Friedmann, N., Belletti, A., & Rizzi, L. (2009). Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua*, 119(1), 67-88. <https://doi.org/10.1016/j.lingua.2008.09.002>
- Friedmann, N., & Novogrodsky, R. (2004). The acquisition of relative clause comprehension in Hebrew: A study of SLI and normal development. *Journal of Child Language*, 31(3), 661-681. <https://doi.org/10.1017/S0305000904006269>

- Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3), e12814. <https://doi.org/10.1111/cogs.12814>
- 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>
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1-76. [https://doi.org/10.1016/S0010-0277\(98\)00034-1](https://doi.org/10.1016/S0010-0277(98)00034-1)
- Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95-126). MIT Press.
- Givón, T. (1983). *Topic continuity in discourse: A quantitative cross-language study*. Benjamins.
- Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6), 1411-1423. <https://doi.org/10.1037/0278-7393.27.6.1411>
- Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2), 274-307. <https://doi.org/10.2307/416535>
- Haendler, Y., Kliegl, R., & Adani, F. (2015). Discourse accessibility constraints in children's processing of object relative clauses. *Frontiers in Psychology*, 6, 860. <https://doi.org/10.3389/fpsyg.2015.00860>
- Hsiao, Y., Dawson, N. J., Banerji, N., & Nation, K. (2022). 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. *Journal of Child Language*, 50(3), 555-580. <https://doi.org/10.1017/S0305000921000957>
- 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. *Language and Cognitive Processes*, 22(6), 860-897. <https://doi.org/10.1080/01690960601155284>

- Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3), 481-507. <https://doi.org/10.1515/COGL.2009.022>
- MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4, 226. <https://doi.org/10.3389/fpsyg.2013.00226>
- Macdonald, R., Brandt, S., Theakston, A., Lieven, E., & Serratrice, L. (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(8), e12874. <https://doi.org/10.1111/cogs.12874>
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates.
- MacWhinney, B., & Pléh, C. (1988). The processing of restrictive relative clauses in Hungarian. *Cognition*, 29(2), 95-141. [https://doi.org/10.1016/0010-0277\(88\)90034-0](https://doi.org/10.1016/0010-0277(88)90034-0)
- Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of Memory and Language*, 47(1), 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>
- Mak, W. M., Vonk, W., & Schriefers, H. (2008). Discourse structure and relative clause processing. *Memory & Cognition*, 36, 170-181. <https://doi.org/10.3758/MC.36.1.170>
- Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Applied Psycholinguistics*, 27(3), 403-422. <https://doi.org/10.1017/S0142716406060334>
- Michaelis, L., & Francis, H. (2007). Lexical subjects and the conflation strategy. In N. Hedberg & R. Zachariski (Eds.), *The grammar-pragmatics interface*. John Benjamins.

- Miyamoto, E., & Nakamura, M. (2003). Subject/Object asymmetries in the processing of relative clauses in Japanese. In G. Garding & M. Tsujimura (Eds.), *Proceedings of WCCFL*, 22, 342-355.
- O'Grady, W., Lee, M., & Choo, M. (2003). A subject-object asymmetry in the acquisition of relative clauses in Korean as a second language. *Studies in Second Language Acquisition*, 25(3), 433-448. <https://doi.org/10.1017/S0272263103000172>
- Pfänder, S. & Behrens, H. (2016). Experience counts: An introduction to frequency effects in language. In H. Behrens & S. Pfänder (Ed.), *Experience Counts: Frequency Effects in Language*(pp. 1-20). De Gruyter. <https://doi.org/10.1515/9783110346916-002>
- Prince, E. F. (1981). Topicalization, focus-movement, and Yiddish-movement: A pragmatic differentiation. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 7, 249-264. <https://doi.org/10.3765/bls.v7i0.2092>
- Randolph, J. J. (2005). *Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa* [Paper presentation]. The Joensuu University Learning and Instruction Symposium, Joensuu, Finland.
- Real, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 57(1), 1-23. <https://doi.org/10.1016/j.jml.2006.08.014>
- Rizzi, L. (1990). *Relativized minimality*. The MIT Press.
- Roland, D., Mauner, G., O'Meara, C., & Yun, H. (2012). Discourse expectations and relative clause processing. *Journal of Memory and Language*, 66(3), 479-508. <https://doi.org/10.1016/j.jml.2011.12.004>
- Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4), 501-532. <https://doi.org/10.1016/j.dr.2007.06.002>
- Song, H. J., & Fisher, C. (2005). Discourse prominence effects on 2.5-year-old children's interpretation of pronouns. *Lingua*, 117(11), 1959-1987. <https://doi.org/10.1016/j.lingua.2006.11.011>
- Tavakolian, S. (1981). The conjoined clause analysis of relative clauses. In S. Tavakolian (Ed.), *Language acquisition and linguistic theory* (pp. 167-187). MIT Press.

Theakston, A. L. (2012). "The spotty cow tickled the pig with a curly tail": How do sentence position, preferred argument structure, and referential complexity affect children's and adult's choice of referring expression?. *Applied Psycholinguistics*, 33(4), 691-724. <https://doi.org/10.1017/S0142716411000531>

Traxler, M. J., Morris, R. K., & Seely, R. E. (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>

Vihman, V. A. (2015). Pick it up: A look at referential devices in Estonian child-directed speech. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 6(2), 63-83. <https://doi.org/10.12697/jeful.2015.6.2.03>

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, 85(1), 79-112. [https://doi.org/10.1016/S0010-0277\(02\)00087-2](https://doi.org/10.1016/S0010-0277(02)00087-2)

### Data, Code and Materials Availability Statement

The present study was pre-registered on the Open Science Framework. All data and analysis scripts are available here:

[https://osf.io/3g68a/overview?view\\_only=9d92bf6efe8f4617916b6a0482a47646](https://osf.io/3g68a/overview?view_only=9d92bf6efe8f4617916b6a0482a47646).

All of our *variables* and the coding scheme (see Appendix A) were pre-registered. The only slight alteration to our pre-registered variables pertains to givenness (see footnote 3). However, after pre-registration, we have developed more specific *predictions* based on the hypotheses stated in the pre-registration. We have also made clear the distinction between an exploratory *technique* and an exploratory *analysis*, which were both referred to as exploratory analysis in the pre-registration. In addition, in the pre-registration, we planned to model our data using linear mixed-effects modelling. Based on reviewer suggestions, we have decided to switch to chi-squared tests that better align with what we set out to test in our hypotheses. The planned mixed-effects modelling is included in the Supplementary Materials.

### Authorship and Contributorship Statement

The authors made the following contributions under the CRediT taxonomy: **Kin Chung Jacky Chan**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing–Original Draft, Writing–Review & Editing, Visualization, Supervision, Project Administration, Funding Acquisition; **Anna Theakston**: Conceptualization, Methodology, Writing–Original Draft, Writing–Review & Editing, Supervision, Funding Acquisition; **Silke Brandt**: Conceptualization, Methodology, Validation, Investigation, Resources, Writing–Original Draft, Writing–Review & Editing, Supervision, Project Administration, Funding Acquisition.

### Acknowledgements

This work was supported by the International Centre for Language and Communicative Development (LuCiD), funded by the Economic and Social Research Council (United Kingdom [ES/S007113/1 and ES/L008955/1]) and the TRINITY student internship programme, funded by the Department of Linguistics and English Language at Lancaster University. We thank Johannes Reilly who supported the coding of corpus data as a research assistant and Rasmus Steinkrauss who provided help on transforming CLAN outputs to accessible Excel format.

### Appendix A

#### Selected corpora

- Thomas Corpus (<https://childes.talkbank.org/access/Eng-UK/Thomas.html>)
  - 2;00.12-3;02.12: One hour, five times a week
  - 3;03.02-3;11.06: One hour (x 5), one week in every month
  - 4;00.02-4;11.20: One hour (x 5), one week in every month
- MPI-EVA-Manchester Corpus (<https://childes.talkbank.org/access/Eng-UK/MPI-EVA-Manchester.html>)
  - Gina
    - 3;00.01-3;01.11: One hour, five times a week, every week
    - 3;02.00-3;11.06: One hour (x 5), one week in every month
    - 4;00.02-4;01.11: One hour, five times a week, every week
    - 4;02.29-4;07.29: One hour (x 5), one week in every month
  - Helen
    - 3;00.01-3;01.11: One hour, five times a week, every week
    - 3;02.00-3;11.10: One hour (x 5), one week in every month
    - 4;00.02-4;01.13: One hour, five times a week, every week
    - 4;02.01-4;11.07: One hour (x 5), one week in every month
    - 5;00.00-5;01.19: One hour, five times a week, every week

#### Procedure for extracting relative clauses from corpora

- Files
  - Child-directed speech: Subset containing the first six weeks starting at age 3;00.01 (five hours per week per child, totalling 60 hours), first week starting at age 4;00.02 (five hours per child, totalling 15 hours), and between 4;02.29 and 4;07.29 (five hours per month per child for five months, totalling 75 hours)

- Child speech: All files in the identified corpora
- Search strings
  - Include utterances that contain any of the words or structures listed below and the 10 immediately preceding utterances (e.g., de Marneffe et al., 2012; Gundel et al., 1993; Michaelis & Francis, 2007; Prince, 1981)
    - The following interrogative/wh- pronouns: who, which, where, whose, and whom
    - A relativizer
    - “that” marked as a complementizer or demonstrative pronoun
    - A noun/pronoun followed by another noun/pronoun followed by a verb/copula/modal verb/auxiliary verb
    - A noun/pronoun followed by a determiner/article followed by a noun
  - Thomas<sup>8</sup>
    - combo +t%mor  
 +s"pro:int|who\*+pro:int|which\*+pro:int|when\*+pro:int|where\*+pro:int|whose\*+pro:int|whom\*+pro:rel|\*+comp|that\*+pro:dem|that\*+((pro\*+n\*)^(pro\*+n\*)^(v|\*+cop|\*+mod|\*+aux\*)))+((pro\*+n\*)^(pro\*aux|\*+n\*aux|\*)))+((pro\*+n\*)^(det:art|\*)^(n\*))" \*.cha +f +u -w10
  - Gina and Helen<sup>1</sup>
    - combo +t%mor  
 +s"pro:wh|who\*+pro:wh|which\*+pro:wh|when\*+pro:wh|where\*+pro:wh|whose\*+pro:wh|whom\*+rel|\*+comp|that\*+pro:dem|that\*+((pro\*+n\*)^(pro\*+n\*)^(v|\*+cop|\*+mod|\*+aux\*)))+((pro\*+n\*)^(pro\*aux|\*+n\*aux|\*)))+((pro\*+n\*)^(det|\*)^(n\*))" \*.cha +f +u -w10
- Decide whether each utterance contains a relative clause
- Copy all target utterances into an excel spreadsheet, number all utterances (ItemNo)
- Exclude utterances from analysis if
  - they are sung nursery rhymes
  - they are incomplete (marked in the transcript by [+ IN]) and syntactically not analysable
  - they do not have a clear head NP (still extract and note the number of such utterances)
- Code each utterance for all variables specified in the coding scheme

---

<sup>8</sup> The search string for Thomas and that for Gina and Helen are essentially the same, the differences were due to the original transcription being done in different versions of CLAN.

**Variable names, variable descriptions, and numeric values (labels)**

| Col | Variable  | Description                                                                                                                                                                                                                                                   | Value | Label             |
|-----|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|-------------------|
| A   | ItemNo    | Running count of target utterances                                                                                                                                                                                                                            |       |                   |
| B   | CHIAge    | Age of child at time of recording (file name)                                                                                                                                                                                                                 |       |                   |
| C   | Speaker   | Indicates whether the target utterance is produced by the mother or the child                                                                                                                                                                                 | 1     | mother            |
|     |           |                                                                                                                                                                                                                                                               | 2     | granny            |
|     |           |                                                                                                                                                                                                                                                               | 3     | child             |
| D   | PreUttRef | Provide a reference for the ease of locating the ten utterances immediately preceding the target utterance – format: OutputFileName_ExcelRowNumber                                                                                                            |       |                   |
| E   | TarUtt    | Target utterance (utterance containing the relative clause) as copied from the transcript<br>Colour-coded: red – Head NP (HNP); blue – Embedded NP(EmNP)<br>If the relative clause has a ditransitive main verb, the direct object will be coded as the EmNP. |       |                   |
| F   | HNP       | Syntactic role of the head NP of the relative clause in the main clause of the target utterance (Diessel & Tomasello, 2000; Fox & Thompson, 1990)                                                                                                             | 1     | subject           |
|     |           |                                                                                                                                                                                                                                                               | 2     | object            |
|     |           |                                                                                                                                                                                                                                                               | 3     | oblique           |
|     |           |                                                                                                                                                                                                                                                               | 4     | predicate nominal |

| Col | Variable           | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Value | Label              |
|-----|--------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|--------------------|
|     |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 5     | existential        |
|     |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 6     | isolated NP        |
| G   | HNPreI             | Syntactic role of the relativized NP within the relative clause in the target utterance (Diessel & Tomasello, 2000; Fox & Thompson, 1990)                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 1     | subject            |
|     |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 2     | direct object      |
|     |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 3     | indirect object    |
|     |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 4     | oblique            |
| H   | Rel                | Relativizer of the relative clause                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |       |                    |
| I   | HNPA <sub>ni</sub> | Indicates animacy of the head NP of the relative clause in the target utterance <ul style="list-style-type: none"> <li>• animate: including (pictures of) humans, animals, body parts, and human and human-like cartoon characters</li> <li>• animized inanimate: including inanimate objects with facial features, toy humans, toy animals, and toy body parts</li> <li>• inanimate: objects that do not move on their own (this includes personified objects, e.g., the police car drove to)</li> <li>• ambiguous: mainly pronouns of which the referent is undiscernible in the preceding 10 utterances</li> </ul> | 1     | animate            |
|     |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 2     | animized inanimate |
|     |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 3     | inanimate          |
|     |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 4     | ambiguous          |

| Col | Variable | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Value | Label              |
|-----|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|--------------------|
| J   | EmNPAni  | Indicates animacy of the embedded NP of the relative clause in the target utterance <ul style="list-style-type: none"> <li>animate: including (pictures of) humans, animals, body parts, and human and human-like cartoon characters</li> <li>animized inanimate: including inanimate objects with facial features, toy humans, toy animals, and toy body parts</li> <li>inanimate: objects that do not move on their own (this includes personified objects)</li> <li>ambiguous: mainly pronouns of which the referent is undiscernible in the preceding 10 utterances</li> </ul> | 1     | animate            |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 2     | animized inanimate |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 3     | inanimate          |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 4     | ambiguous          |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 5     | N/A                |
| K   | HNPTtype | Type of the head NP of the relative clause in the target utterance                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 1     | pronoun            |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 2     | proper noun        |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 3     | lexical NP         |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 4     | other              |

| Col | Variable   | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Value | Label            |
|-----|------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|------------------|
| L   | HNPProType | <p>For head NPs being proper nouns, indicates proper noun type</p> <ul style="list-style-type: none"> <li>child's own name: if the head NP is the child's own name or variation of the child's own name</li> <li>high frequency: if the head NP is frequent in the corpus (i.e., accounting for more than 1% of the proper noun tokens in each corpus), the following lists specify the words that are considered high frequency (excluding the child's own name) in each corpus <ul style="list-style-type: none"> <li>Thomas: Mummy; Purdie; Daddy; Christmas; Grandma; Grandad; Po; Postman_Pat</li> <li>Gina: Mummy; Grandma; Nanny; Madam; Daddy; Jwww; Gwww</li> <li>Helen: Mummy; Daddy; Jwww; Awww; Fifi; Cwww; Dora; Ewww; Lwww; Nwww</li> </ul> </li> <li>general: if otherwise</li> </ul> | 1     | child's own name |
|     |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 2     | high frequency   |
|     |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 3     | general          |
|     |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 4     | N/A              |
| M   | HNPPer     | Person and number of the head NP of the relative clause in the target utterance                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 1     | 1SG              |
|     |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 2     | 2SG              |
|     |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 3     | 3SG              |
|     |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 4     | 1PL              |
|     |            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 5     | 2PL              |

| Col | Variable   | Description                                                                                       | Value | Label         |
|-----|------------|---------------------------------------------------------------------------------------------------|-------|---------------|
|     |            |                                                                                                   | 6     | 3PL           |
|     |            |                                                                                                   | 7     | unclear       |
|     |            |                                                                                                   | 8     | N/A           |
| N   | HNPPref    | For 3SG pronoun head NPs, indicates whether the referent is a real-world referent or an expletive | 1     | real          |
|     |            |                                                                                                   | 2     | expletive     |
|     |            |                                                                                                   | 3     | N/A           |
| O   | HNPLexType | For lexical head NPs, indicates whether it is definite, indefinite, bare, a gerund, or other      | 1     | definite NP   |
|     |            |                                                                                                   | 2     | indefinite NP |
|     |            |                                                                                                   | 3     | bare NP       |
|     |            |                                                                                                   | 4     | gerund        |
|     |            |                                                                                                   | 5     | other         |
|     |            |                                                                                                   | 6     | N/A           |
| P   | EmNPType   | Type of the embedded NP of the relative clause in the target utterance                            | 1     | pronoun       |
|     |            |                                                                                                   | 2     | proper noun   |
|     |            |                                                                                                   | 3     | lexical NP    |
|     |            |                                                                                                   | 4     | other         |
|     |            |                                                                                                   | 5     | N/A           |

| Col | Variable    | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Value | Label            |
|-----|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|------------------|
| Q   | EmNPProType | <p>For embedded NPs being proper nouns, indicates proper noun type</p> <ul style="list-style-type: none"> <li>child's own name: if the embedded NP is the child's own name or variation of the child's own name</li> <li>high frequency: if the embedded NP is frequent in the corpus (i.e., accounting for more than 1% of the proper noun tokens in each corpus), the following lists specify the words that are considered high frequency (excluding the child's own name) in each corpus <ul style="list-style-type: none"> <li>Thomas: Mummy; Purdie; Daddy; Christmas; Grandma; Grandad; Po; Postman_Pat</li> <li>Gina: Mummy; Grandma; Nanny; Madam; Daddy; Jwww; Gwww</li> <li>Helen: Mummy; Daddy; Jwww; Awww; Fifi; Cwww; Dora; Ewww; Lwww; Nwww</li> </ul> </li> <li>general: if otherwise</li> </ul> | 1     | child's own name |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 2     | high frequency   |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 3     | general          |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 4     | N/A              |
| R   | EmNPPer     | Person and number of the embedded NP of the relative clause in the target utterance                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | 1     | 1SG              |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 2     | 2SG              |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 3     | 3SG              |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 4     | 1PL              |

| Col | Variable    | Description                                                                                           | Value | Label         |
|-----|-------------|-------------------------------------------------------------------------------------------------------|-------|---------------|
|     |             |                                                                                                       | 5     | 2PL           |
|     |             |                                                                                                       | 6     | 3PL           |
|     |             |                                                                                                       | 7     | unclear       |
|     |             |                                                                                                       | 8     | N/A           |
| S   | EmNPRef     | For 3SG pronoun embedded NPs, indicates whether the referent is a real-world referent or an expletive | 1     | real          |
|     |             |                                                                                                       | 2     | expletive     |
|     |             |                                                                                                       | 3     | N/A           |
| T   | EmNPLexType | For lexical embedded NPs, indicates whether it is definite, indefinite, bare, a gerund, or other      | 1     | definite NP   |
|     |             |                                                                                                       | 2     | indefinite NP |
|     |             |                                                                                                       | 3     | bare NP       |
|     |             |                                                                                                       | 4     | gerund        |
|     |             |                                                                                                       | 5     | other         |
|     |             |                                                                                                       | 6     | N/A           |

| Col | Variable    | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Value | Label      |
|-----|-------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|------------|
| U   | HNPGiv      | <p>Givenness (referential and lexical) of the head NP of the relative clause in the target utterance</p> <ul style="list-style-type: none"> <li>• given: if the head NP is mentioned or referred to in the preceding 10 utterances (even if it refers to a different referent), including the use of pronoun to replace a lexical NP, or clearly predictable from the context (e.g., both the mother and the child are attending to it; Roland et al., 2012), or if the referent is present in the visual scene</li> <li>• associated: if the embedded NP is semantically associated with a referent, action, or concept in the preceding 10 utterances (e.g., ladies -&gt; people)</li> <li>• new: if otherwise (but code as unclear if any of the 10 previous utterances are incomplete or transcribed as 'xxx' or 'www')</li> </ul> | 1     | given      |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 2     | associated |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 3     | new        |
|     |             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 4     | unclear    |
| V   | HNPRelation | What the given/associated head NP relates to and where                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |       |            |

| Col | Variable     | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | Value | Label      |
|-----|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|------------|
| W   | EmNPGiv      | <p>Givenness (referential and lexical) of the embedded NP of the relative clause in the target utterance</p> <ul style="list-style-type: none"> <li>• given: if the embedded NP is mentioned or referred to in the preceding 10 utterances (even if it refers to a different referent), including the use of pronoun to replace a lexical NP, or clearly predictable from the context (e.g., both the mother and the child are attending to it; Roland et al., 2012), or if the referent is present in the visual scene</li> <li>• associated: if the embedded NP is semantically associated with a referent, action, or concept in the preceding 10 utterances (e.g., ladies -&gt; people)</li> <li>• new: if otherwise (but code as unclear if any of the 10 previous utterances are incomplete or transcribed as 'xxx' or 'www')</li> </ul> | 1     | given      |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 2     | associated |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 3     | new        |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 4     | unclear    |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | 5     | N/A        |
| X   | EmNPRelation | What the given/associated embedded NP relates to and where                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |       |            |

| Col | Variable | Description                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Value | Label     |
|-----|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|-----------|
| Y   | HNPTop   | <p>Topichood of the head NP of the relative clause in the target utterance</p> <ul style="list-style-type: none"> <li>• topic: if the head NP is the topic of the ongoing discourse (i.e., focus of conversation) in the preceding 10 utterances; if the topic changes during the preceding 10 utterances, the topic closest to the target utterance is considered</li> <li>• not topic: if otherwise</li> </ul> <p>But</p> <ul style="list-style-type: none"> <li>• unclear: if any of the 10 previous utterances are incomplete or transcribed as 'xxx' or 'www'</li> </ul>         | 1     | topic     |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 2     | non-topic |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 3     | unclear   |
| Z   | EmNPTop  | <p>Topichood of the embedded NP of the relative clause in the target utterance</p> <ul style="list-style-type: none"> <li>• topic: if the embedded NP is the topic of the ongoing discourse (i.e., focus of conversation) in the preceding 10 utterances; if the topic changes during the preceding 10 utterances, the topic closest to the target utterance is considered</li> <li>• not topic: if otherwise</li> </ul> <p>But</p> <ul style="list-style-type: none"> <li>• unclear: if any of the 10 previous utterances are incomplete or transcribed as 'xxx' or 'www'</li> </ul> | 1     | topic     |
|     |          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 2     | non-topic |

| Col | Variable    | Description                                                                                                                            | Value | Label                |
|-----|-------------|----------------------------------------------------------------------------------------------------------------------------------------|-------|----------------------|
|     |             |                                                                                                                                        | 3     | unclear              |
|     |             |                                                                                                                                        | 4     | N/A                  |
| AA  | EmVerb      | Main / phrasal verb (in base form; i.e., 'is', 'am', and 'are' should be coded as 'be') of the relative clause in the target utterance |       |                      |
| AB  | EmVerbType  | Type of the verb of the relative clause (according to construction) in the target utterance                                            | 1     | intransitive         |
|     |             |                                                                                                                                        | 2     | monotransitive       |
|     |             |                                                                                                                                        | 3     | ditransitive         |
|     |             |                                                                                                                                        | 4     | copula               |
|     |             |                                                                                                                                        | 5     | existential          |
|     |             |                                                                                                                                        | 6     | passive              |
|     |             |                                                                                                                                        | 7     | other                |
| AC  | EmVerbArgOb | For intransitive embedded verb, states whether the argument is obligatory<br>Use N/A if no embedded NP                                 | 1     | obligatory location  |
|     |             |                                                                                                                                        | 2     | obligatory recipient |
|     |             |                                                                                                                                        | 3     | non-obligatory       |
|     |             |                                                                                                                                        | 4     | unclear              |
|     |             |                                                                                                                                        | 5     | N/A                  |
| AD  | EmVerbPass  | For passives, states whether it is with or without agent                                                                               | 1     | without agent        |
|     |             |                                                                                                                                        | 2     | with agent           |

| Col | Variable  | Description                                                                                                                                                                                                                                                                                                                                                                                                                             | Value | Label         |
|-----|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|---------------|
|     |           |                                                                                                                                                                                                                                                                                                                                                                                                                                         | 3     | N/A           |
| AE  | EmNPTypeV | NP type and verb combination (for generating frequency data)                                                                                                                                                                                                                                                                                                                                                                            |       |               |
| AF  | EmNPVType | NP type and verb type combination (for generating frequency data)                                                                                                                                                                                                                                                                                                                                                                       |       |               |
| AG  | Ellipsis  | Indicates whether the target utterance contains an ellipsis where the elided information is recoverable from the linguistic context – this is often expressed by ‘xxx’ <ul style="list-style-type: none"> <li>no ellipsis: all structures and information that need to be coded are present and clear – no need to recover anything (even if there is xxx marker elsewhere in the utterance)</li> <li>ellipsis: if otherwise</li> </ul> | 1     | no ellipsis   |
|     |           |                                                                                                                                                                                                                                                                                                                                                                                                                                         | 2     | ellipsis      |
| AH  | UttType   | Indicates whether the target utterance is declarative, imperative, or interrogative<br>Interrogative includes <i>wh</i> - questions, yes/no questions, and prosodic questions (utterances that are syntactically declarative but with prosody that signals a question; i.e., utterance ending with a question mark)<br>Tag questions are coded as declarative, as tag questions often do not function as questions                      | 1     | declarative   |
|     |           |                                                                                                                                                                                                                                                                                                                                                                                                                                         | 2     | imperative    |
|     |           |                                                                                                                                                                                                                                                                                                                                                                                                                                         | 3     | interrogative |

| Col | Variable     | Description                                                                                                                                                                                                                                                                                                                                                                                                            | Value | Label        |
|-----|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|--------------|
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                        | 4     | unclear      |
| AI  | PrevUttQ     | Indicates whether the utterance immediately before the target utterance is a question or not<br>Question includes <i>wh</i> - questions, yes/no questions, and prosodic questions (utterances that are syntactically declarative but with prosody that signals a question; i.e., utterance ending with a question mark)<br>Tag questions are coded as declarative, as tag questions often do not function as questions | 1     | question     |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                        | 2     | not question |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                        | 3     | unclear      |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                        | 4     | N/A          |
| AJ  | PrevUttQWord | If the utterance immediately before the target utterance is a question, the <i>wh</i> - word or auxiliary verb used                                                                                                                                                                                                                                                                                                    |       |              |
| AK  | Repeat       | Indicates whether the target utterance is a repetition of a previous utterance in the 10 utterances immediately before the target utterance <ul style="list-style-type: none"> <li>• repetition: if it is a verbatim repetition</li> <li>• recast: if any changes occur in an attempt to repeat a child's utterance</li> <li>• independent: if otherwise (an independent new contribution)</li> </ul>                  | 1     | repetition   |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                        | 2     | recast       |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                        | 3     | independent  |
|     |              |                                                                                                                                                                                                                                                                                                                                                                                                                        | 4     | unclear      |

### Inter-rater reliability of coding

- Number of relative clauses second-coded for reliability check by corpus and speaker

| Corpus | Speaker   | No. of RCs | No. of second-coded RCs |
|--------|-----------|------------|-------------------------|
| Gina   | Caregiver | 227        | 36 (15.86%)             |
|        | Child     | 79         | 14 (17.72%)             |
| Helen  | Caregiver | 262        | 53 (20.23%)             |
|        | Child     | 276        | 57 (20.65%)             |
| Thomas | Caregiver | 737        | 148 (20.08%)            |
|        | Child     | 140        | 27 (19.29%)             |

- Free-marginal multirater kappa ( $\kappa$ ) of each categorical variable by corpus and speaker

| Corpus | Speaker   | HNP   | HNPRel | HNPA <sub>ni</sub> | EmNPA <sub>ni</sub> | HNPT <sub>type</sub> | HNPPro <sub>type</sub> |
|--------|-----------|-------|--------|--------------------|---------------------|----------------------|------------------------|
| Gina   | Caregiver | .967  | 1.000  | .926               | .965                | .926                 | 1.000                  |
|        | Child     | 1.000 | 1.000  | .905               | 1.000               | 1.000                | 1.000                  |
| Helen  | Caregiver | .932  | .975   | .950               | .929                | .950                 | 1.000                  |
|        | Child     | .874  | .977   | .743               | .912                | .953                 | 1.000                  |
| Thomas | Caregiver | .984  | 1.000  | .991               | .966                | .982                 | 1.000                  |
|        | Child     | .867  | .951   | .901               | .954                | .852                 | 1.000                  |

| Corpus | Speaker   | HNPPer | HNPRef | HNPLexType | EmNPType | EmNPProType | EmNPPer |
|--------|-----------|--------|--------|------------|----------|-------------|---------|
| Gina   | Caregiver | 1.000  | .917   | .933       | .965     | 1.000       | .968    |
|        | Child     | 1.000  | 1.000  | 1.000      | 1.000    | 1.000       | 1.000   |
| Helen  | Caregiver | 1.000  | .943   | .932       | .953     | 1.000       | .957    |
|        | Child     | .920   | .895   | .958       | .978     | .977        | .960    |
| Thomas | Caregiver | 1.000  | .990   | .968       | .975     | .973        | .954    |
|        | Child     | .958   | .944   | .956       | 1.000    | 1.000       | 1.000   |

| Corpus | Speaker   | EmNPRef | EmNPLeType | HNPGiv | EmNPGiv | HNPTop | EmNPPTop |
|--------|-----------|---------|------------|--------|---------|--------|----------|
| Gina   | Caregiver | 1.000   | .967       | .481   | .896    | .750   | .778     |
|        | Child     | .893    | 1.000      | .524   | .732    | .893   | .905     |
| Helen  | Caregiver | .972    | .909       | .950   | .858    | .915   | .899     |
|        | Child     | .974    | .958       | .696   | .803    | .816   | .836     |
| Thomas | Caregiver | 1.000   | .976       | .928   | .975    | .949   | .982     |
|        | Child     | 1.000   | 1.000      | .753   | .907    | .722   | .852     |

| Corpus | Speaker   | EmVerbType | EmVerbArg | EmVerbPass | Ellipsis | UttType | PreUttQ |
|--------|-----------|------------|-----------|------------|----------|---------|---------|
| Gina   | Caregiver | .773       | .757      | 1.000      | 1.000    | 1.000   | .963    |
|        | Child     | 1.000      | 1.000     | 1.000      | 1.000    | .905    | .905    |
| Helen  | Caregiver | .890       | .882      | 1.000      | 1.000    | .975    | 1.000   |
|        | Child     | .918       | .912      | 1.000      | 1.000    | .930    | .883    |
| Thomas | Caregiver | .945       | .966      | .990       | 1.000    | 1.000   | .982    |
|        | Child     | .957       | .954      | 1.000      | .852     | .951    | .901    |

| Corpus | Speaker   | Repeat |
|--------|-----------|--------|
| Gina   | Caregiver | .926   |
|        | Child     | .429   |
| Helen  | Caregiver | .950   |
|        | Child     | .906   |
| Thomas | Caregiver | .973   |
|        | Child     | .852   |

## Appendix B

**Table B.1. Descriptive statistics for different types of NPs by corpus**

|                        | Child-directed speech |       |        |           |       |        | Child speech |       |        |           |       |        |
|------------------------|-----------------------|-------|--------|-----------|-------|--------|--------------|-------|--------|-----------|-------|--------|
|                        | Subject RC            |       |        | Object RC |       |        | Subject RC   |       |        | Object RC |       |        |
|                        | Gina                  | Helen | Thomas | Gina      | Helen | Thomas | Gina         | Helen | Thomas | Gina      | Helen | Thomas |
| Head NP animate        | 29                    | 58    | 100    | 9         | 16    | 25     | 26           | 51    | 19     | 1         | 12    | 5      |
| Head NP inanimate      | 35                    | 47    | 131    | 121       | 102   | 288    | 8            | 41    | 26     | 30        | 92    | 52     |
| Embedded NP animate    | 11                    | 22    | 40     | 128       | 119   | 327    | 9            | 21    | 7      | 33        | 116   | 56     |
| Embedded NP inanimate  | 34                    | 48    | 124    | 3         | 1     | 6      | 13           | 41    | 19     | 0         | 0     | 2      |
| Head NP pronoun        | 11                    | 16    | 28     | 16        | 6     | 41     | 4            | 21    | 7      | 6         | 19    | 21     |
| Head NP lexical NP     | 55                    | 84    | 204    | 114       | 113   | 287    | 31           | 88    | 42     | 27        | 99    | 39     |
| Embedded NP pronoun    | 14                    | 13    | 29     | 120       | 111   | 278    | 6            | 9     | 10     | 33        | 115   | 52     |
| Embedded NP lexical NP | 33                    | 55    | 121    | 1         | 2     | 9      | 16           | 49    | 19     | 0         | 1     | 5      |
| Head NP given          | 35                    | 63    | 125    | 56        | 72    | 193    | 24           | 69    | 23     | 22        | 73    | 19     |
| Head NP new            | 18                    | 29    | 68     | 35        | 25    | 84     | 5            | 29    | 15     | 5         | 26    | 19     |
| Embedded NP given      | 26                    | 25    | 76     | 119       | 113   | 298    | 14           | 24    | 12     | 33        | 112   | 51     |
| Embedded NP new        | 13                    | 25    | 63     | 4         | 5     | 25     | 4            | 30    | 10     | 0         | 4     | 4      |
| Head NP topic          | 32                    | 44    | 76     | 53        | 48    | 123    | 14           | 35    | 11     | 14        | 44    | 12     |
| Head NP non-topic      | 32                    | 53    | 151    | 62        | 57    | 179    | 16           | 74    | 32     | 16        | 62    | 41     |
| Embedded NP topic      | 14                    | 7     | 15     | 8         | 2     | 10     | 4            | 3     | 4      | 1         | 9     | 1      |
| Embedded NP non-topic  | 32                    | 59    | 151    | 114       | 113   | 305    | 15           | 63    | 23     | 29        | 101   | 55     |

*Note.* Cells shaded in grey indicate relative percentages of subject RC and object RC showing a different pattern to the general pattern when looking across all corpora.

## **Supplementary Materials**

### **Deviation from Pre-Registration**

In the pre-registration, we planned to include child age (in months) to model child data. However, this was not possible as the number of RCs identified in our corpora was small and did not justify models including an extra variable of age and the interactions between age and all other variables. Another deviation from our pre-registration was that we used a more powerful parametric approach – generalized linear mixed-effects modelling – to examine differences identified between child-directed and child speech instead of the planned non-parametric chi-squared tests.

### **Preliminary Analyses**

Here, we provide a short summary of our preliminary analyses using exploratory analytical techniques. Full details of these analyses can be found on OSF (see link in the Data Availability section). The main purpose of these preliminary analyses was to explore whether RCs produced by children and their caregivers contain the same features as the RCs found in adult-directed speech and written corpora (Fox & Thompson, 1990; Roland et al., 2012). Some of these features, such as definiteness of head and embedded NPs, have not been explored in any previous studies investigating children's and caregivers' production of RCs. In addition, our descriptive statistics summarized above already indicate that the RCs produced by children and caregivers do not always show the same discourse-based features as the ones identified in adult-directed speech and written corpora.

Our random forest analyses reveal that the following attributes of embedded NPs differ between subject and object RCs: animacy, type, person and number, type of lexical NP, givenness, and topichood. For example, compared to subject RCs, object RCs are more likely to contain an animate embedded NP (see also Table 3). In addition, animacy and person and number of the head NP differ between subject and object RCs. For example, compared to subject RCs, object RCs are more likely to occur with an inanimate head NP (see also Table 3). These variables consistently rank highly in our analyses across all corpora (for both child and caregiver data). Our MCAs also suggest that type, definiteness (i.e., type of lexical NP), givenness, and topichood of the head NP may differ between subject and object RCs. We included these variables as fixed effects in our GLMs to test whether they significantly differ between subject and object RCs. In addition, the splits in the conditional inference tree analyses suggest that the following variables may interact to characterize type of RC: (1) animacy of embedded NP and animacy of head NP, (2) animacy of embedded NP and type of embedded lexical NP, (3) type of embedded NP and animacy of head NP, and (4) type and animacy of embedded NP. We therefore included these as interaction terms in our GLMs.

## Generalized Linear Mixed-Effects Modelling<sup>9</sup>

We ran separate models for the child and caregiver data to determine which predictors, identified in our preliminary analyses, best characterize whether a RC is a subject or object RC (e.g., whether the animacy status of the head NP can predict whether the following RC is a subject or object RC).

### Caregiver Data

The caregiver data for the GLMs consisted of 1008 observations (subject and object RCs). A series of GLMs were fitted using the `glmer` function (family = binomial) in the `lme4` package in R. Subject RCs were coded as 0 and object RCs as 1. In all models, all predictors were entered simultaneously. First, assuming the same random effects of corpus on intercepts, the following models differing in fixed effects were compared:

- Model 1      A model with just the intercept.
- Model 2      The final model was obtained through the following backwards elimination steps: (a) A saturated model with all identified main effects and interaction terms as specified above was first fitted to the data. (b) Then, the `drop1` function (test = “Chisq”) was used to determine whether dropping the highest order fixed effect with the highest *p*-value from the model would fit the data better. (c) The highest order fixed effect with the highest likelihood ratio test *p*-value once dropped was then removed from the model, and a model with the identified fixed effect removed was then fitted to the data. (d) Steps (b) to (c) were then repeated until all likelihood ratio test *p*-values between a more complex model and all simpler models with one of the highest order fixed effects removed from the complex model was smaller than .05 to obtain the final model.<sup>10</sup>

---

<sup>9</sup> To ensure the robustness of the mixed-effects modelling results with only three levels to the random effect of Corpus, we have conducted alternative regression analyses without random effects and with Corpus as a fixed effect and included its interaction with other variables (with the same backwards elimination approach described in this section). We obtained final models with the same result patterns. The fixed effects involving Corpus were all dropped in the elimination process. We also ran a series of models where the main effect of Corpus was kept in the model throughout the elimination process and obtained the same result patterns, with the main effect of Corpus not reaching statistical significance. We report the mixed-effects models here because we are not primarily interested in differences between the three corpora, but rather, how well any observed patterns can be generalized when taking these potential differences into account.

<sup>10</sup> Reference level of each factor was set to the level with value 1 in the coding scheme (see Appendix A). This approach captures all interesting and most meaningful comparisons

Model 2 was the preferred model: comparing Models 1 and 2,  $\chi^2(19) = 1062.50$ ,  $p < .001$ . Model 2 was then compared to the most preferred model (based on likelihood ratio tests) that included random effects of corpus on the slopes of the fixed effects (Model 3). We found that the inclusion of random effects of corpus on the slopes of the fixed effects was not justified,  $\chi^2(9) = 1.35$ ,  $p = .998$ . As a result, Model 2 was selected as the final model (AIC = 349, BIC = 452, logLik = -154, deviance = 307). A summary of Model 2 is reported in Table S1.

Based on Model 2, the following fixed effects significantly predicted whether a RC is subject or object RC: animacy of the head NP, and animacy, type, and topichood of the embedded NP. Compared to subject RCs, object RCs were more likely to contain an inanimate head NP (H3). At the same time, subject RCs tended to be used with inanimate embedded NPs. These results suggest that both object and subject RCs tend to contain inanimate objects / patients. Subject RCs also tended to be used with embedded NPs expressed by proper nouns rather than pronouns. Finally, contradicting H2, object RCs were more likely to be used with embedded NPs that were not the topic of the ongoing conversation.

### **Child Data**

The child data for GLM consisted of 412 observations (subject and object RCs). Using the same approach as for caregiver data, all fixed effects were entered simultaneously in all models. The following models which assumed the same random effects of corpus on intercepts but differed in fixed effects were compared:

Model 4      A model with just the intercept.

Model 5      The final model obtained through similar steps of obtaining Model 2.

---

(e.g., given against new for givenness), except for person and number, for which the most meaningful comparison was less obvious. Additional sets of GLMs were run with other levels of person and number set as the reference level. The final models for these additional sets of GLMs were the same as the one reported here.

**Table S1. Summary of Model 2, modelled on caregiver data: HNPA<sub>ni</sub> + EmNPA<sub>ni</sub> + EmNPType + EmNPLe<sub>x</sub>Type + EmNPTo<sub>p</sub> + (1|Corpus)**

| Fixed effects                       | Estimated coefficient | SE           | Wald confidence intervals |               | <i>z</i>      | <i>pr(&gt; z )</i> |
|-------------------------------------|-----------------------|--------------|---------------------------|---------------|---------------|--------------------|
|                                     |                       |              | 2.50%                     | 97.50%        |               |                    |
| (Intercept)                         | -22.154               | 5925.477     | -1.164e+04                | 1.159e+04     | -0.004        | .9970              |
| HNPA <sub>ni</sub> (animate)        |                       |              |                           |               |               |                    |
| animized inanimate                  | 0.716                 | 0.858        | -9.659e-01                | 2.397         | 0.834         | .4042              |
| <b>inanimate</b>                    | <b>2.845</b>          | <b>0.355</b> | <b>2.150</b>              | <b>3.540</b>  | <b>8.024</b>  | <b>&lt; .0001</b>  |
| ambiguous                           | 1.779                 | 1.012        | -2.039                    | 3.762         | 1.758         | .0787              |
| EmNPA <sub>ni</sub> (animate)       |                       |              |                           |               |               |                    |
| animized inanimate                  | -2.022                | 2.119        | -6.174                    | 2.130         | -0.955        | .3398              |
| <b>inanimate</b>                    | <b>-4.125</b>         | <b>0.470</b> | <b>-5.047</b>             | <b>-3.203</b> | <b>-8.770</b> | <b>&lt; .0001</b>  |
| ambiguous                           | -22.746               | 6864.846     | -1.348e+04                | 1.343e+04     | -0.003        | .9974              |
| non-applicable                      | 10.127                | 9294.152     | -1.821e+04                | 1.823e+04     | 0.001         | .9991              |
| EmNPType (pronoun)                  |                       |              |                           |               |               |                    |
| <b>proper noun</b>                  | <b>-0.880</b>         | <b>0.442</b> | <b>-1.746</b>             | <b>-1.356</b> | <b>-1.991</b> | <b>.0465</b>       |
| lexical NP                          | 17.914                | 5925.477     | -1.160e+04                | 1.163e+04     | 0.003         | .9976              |
| other                               | -1.350                | 1.037        | -3.382                    | 6.828e-01     | -1.302        | .1931              |
| non-applicable                      | -32.757               | 10162.798    | -1.995e+04                | 1.989e+04     | -0.003        | .9974              |
| EmNPLe <sub>x</sub> Type (definite) |                       |              |                           |               |               |                    |
| indefinite                          | -0.241                | 0.930        | -2.063                    | 1.581         | -0.259        | .7956              |
| bare                                | -16.107               | 4713.290     | -9.254e+03                | 9.222e+03     | -0.003        | .9973              |
| gerund                              | -11.200               | 17730.378    | -3.476e+04                | 3.474e+04     | -0.001        | .9995              |
| other                               | 38.722                | 18694.309    | -3.660e+04                | 3.668e+04     | 0.002         | .9984              |
| non-applicable                      | 21.728                | 5925.477     | -1.159e+04                | 1.164e+04     | 0.004         | .9971              |
| EmNPTo <sub>p</sub> (topic)         |                       |              |                           |               |               |                    |
| <b>non-topic</b>                    | <b>1.503</b>          | <b>0.522</b> | <b>4.789e-01</b>          | <b>2.526</b>  | <b>2.877</b>  | <b>.0040</b>       |
| unclear                             | 1.330                 | 0.826        | -2.891e-01                | 2.948         | 1.610         | .1074              |
| non-applicable                      | 15.893                | 4111.107     | -8.042e+03                | 8.074e+03     | 0.004         | .9969              |
| Random effects                      | Name                  | Variance     | SD                        |               |               |                    |
| Subject                             | (Intercept)           | 5.197e-15    | 7.209e-08                 |               |               |                    |
|                                     | AIC                   | BIC          | logLik                    | Deviance      |               |                    |
|                                     | 349                   | 452          | -154                      | 307           |               |                    |

*Notes.* Reference level of each main effect is given in parentheses. A positive effect indicates a trend for object RC, whereas a negative effect indicates a trend for subject RC.

Model 5 was the preferred model: comparing Models 4 and 5,  $\chi^2(18) = 489.14$ ,  $p < .001$ . Then, Model 5 was compared to the most preferred model (based on likelihood ratio tests) that consisted of random effects of corpus on the slopes of the fixed effects (Model 6). We found that the inclusion of random effects of corpus on the slopes of the fixed effects was not justified,  $\chi^2(14) = 4.23$ ,  $p = .994$ . Thus, Model 5 was selected as the final model (AIC = 122, BIC = 202, logLik = -41, deviance = 82). A summary of Model 5 is reported in Table S2.

Animacy and type of lexical NP of the head NP, and animacy, type, and topichood of the embedded NP significantly predicted type of RC. Similar to the pattern observed in the caregiver data, object RCs were more likely to be used with inanimate head NPs (H3). And subject RCs tended to contain inanimate embedded NPs or embedded NPs that were ambiguous in terms of animacy status. Unlike caregivers, children also showed a significant tendency to produce subject RCs with indefinite head NPs (H5). In addition, the embedded NP of children's subject RCs tended to be expressed by proper or lexical nouns rather than pronouns, which partly reflect the caregiver data again. Also similar to the pattern found in the caregiver data, the embedded NP of children's object RCs was unlikely to be the ongoing topic of the conversation. Instead, children's object RCs tended to contain embedded NPs that could not clearly be coded as topic or non-topic. We will come back to this unexpected finding and how it might be related to our coding scheme in the Discussion section.

### ***Comparing Caregiver and Child Data***

Next, to more closely look at how the lexical, semantic, and discourse-level characteristics of subject and object RCs in the children's input align with those in their own production, we conducted GLMs on the complete dataset consisting of both the caregiver and child data. For each of the main effects that predict differently in Models 2 and 5 (i.e., main effects containing any meaningful categories with at least marginally significant effects going in different directions), we ran a separate GLM. Therefore, we ran separate models for: type of lexical head NP, and type, givenness, and topichood of embedded NP. These GLMs each included one of the identified main effects, an additional factor of speaker group (i.e., whether the RC is produced by a child or a caregiver), and the two-way interaction term of this additional factor of speaker and the identified main effect.

The data for these GLMs consisted of 1420 observations (subject and object RCs). All fixed effects were entered simultaneously into the models. We included speaker identity as the random intercept. Note that here we did not adopt the same model comparison approach as outlined above, as we were interested in qualifying the differences found between the patterns observed in the caregiver data and the child data. We did not include random slopes in these models, as these were not justified in Models 2 and 5.

|                       |               |              | Wald confidence  |                   |               |                  |
|-----------------------|---------------|--------------|------------------|-------------------|---------------|------------------|
|                       | Estimated     |              | intervals        |                   |               |                  |
| Fixed effects         | coefficient   | SE           | 2.50%            | 97.50%            | z             | pr(> z )         |
| (Intercept)           | 1.096         | 1.348        | -1.546           | 3.738             | 0.813         | .4161            |
| HNPNani (animate)     |               |              |                  |                   |               |                  |
| animized inanimate    | 4.713         | 5.157        | -5.395           | 1.482e+01         | 0.914         | .3608            |
| <b>inanimate</b>      | <b>5.743</b>  | <b>1.062</b> | <b>3.661</b>     | <b>7.824</b>      | <b>5.408</b>  | <b>&lt;.0001</b> |
| <b>ambiguous</b>      | <b>1.734</b>  | <b>0.874</b> | <b>2.061e-02</b> | <b>3.446</b>      | <b>1.984</b>  | <b>.0473</b>     |
| HNPLexType (definite) |               |              |                  |                   |               |                  |
| <b>indefinite</b>     | <b>-2.349</b> | <b>0.910</b> | <b>-4.132</b>    | <b>-5.67e+01</b>  | <b>-2.583</b> | <b>.0098</b>     |
| bare                  | 3.504         | 46.891       | -8.840e+01       | 9.541e+01         | 0.075         | .9404            |
| non-applicable        | 0.808         | 0.858        | -8.737e-01       | 2.489             | 0.941         | .3465            |
| EmNPNani (animate)    |               |              |                  |                   |               |                  |
| animized inanimate    | -16.679       | 79462.015    | -1.558e+05       | 1557e+05          | 0.000         | .9998            |
| <b>inanimate</b>      | <b>-6.329</b> | <b>1.321</b> | <b>-8.917</b>    | <b>-3.740</b>     | <b>-4.792</b> | <b>&lt;.0001</b> |
| <b>ambiguous</b>      | <b>-5.012</b> | <b>1.616</b> | <b>-8.180</b>    | <b>-1.844</b>     | <b>-3.101</b> | <b>.0019</b>     |
| non-applicable        | -27.234       | 34717.485    | -6.807e+04       | 6.802e+04         | -0.001        | .9994            |
| EmNPType (pronoun)    |               |              |                  |                   |               |                  |
| <b>proper noun</b>    | <b>-4.581</b> | <b>1.867</b> | <b>-8.240</b>    | <b>-9.212e-01</b> | <b>-2.453</b> | <b>.0142</b>     |
| <b>lexical NP</b>     | <b>-6.044</b> | <b>1.527</b> | <b>-9.037</b>    | <b>-3.052</b>     | <b>-3.958</b> | <b>&lt;.0001</b> |
| other                 | -29.909       | 42799.044    | -8.391e+04       | 8.385e+04         | -0.001        | .9994            |
| EmNPNGiv (given)      |               |              |                  |                   |               |                  |
| new                   | 2.562         | 1.505        | -3.876e-01       | 5.511             | 1.702         | .0887            |
| unclear               | -1.132        | 1.377        | -3.830           | 1.567             | -0.822        | .4111            |
| EmNPTop (topic)       |               |              |                  |                   |               |                  |
| non-topic             | -0.615        | 1.354        | -3.269           | 2.039             | -0.454        | .6496            |
| <b>unclear</b>        | <b>4.091</b>  | <b>1.877</b> | <b>4.136e-01</b> | <b>7.769</b>      | <b>2.180</b>  | <b>.0292</b>     |
| non-applicable        | -1.025        | 35481.103    | -6.954e+04       | 6.954e+04         | 0.000         | 1.0000           |
|                       |               |              |                  |                   |               |                  |
| Random effects        | Name          | Variance     | SD               |                   |               |                  |
| Subject               | (Intercept)   | 1.288e-16    | 1.135e-08        |                   |               |                  |
|                       |               |              |                  |                   |               |                  |
|                       | AIC           | BIC          | logLik           | Deviance          |               |                  |
|                       | 122           | 202          | -41              | 82                |               |                  |

Volume 6, Issue 1

A significant interaction with speaker group was only found in the final models for type of lexical head NP and type of embedded NP. For lexical head NP, this interaction shows that caregivers', but not children's, subject RCs tended to be used without lexical head NPs (i.e., with other types of NPs as the head NP) ( $\beta = 0.718$ ,  $SE = 0.315$ ,  $p = .0229$ , 95%CI [0.100, 1.336]). For type of embedded NP, this interaction shows that children's, but not caregivers', subject RCs were likely to contain embedded NPs expressed by proper nouns (compared to pronouns) ( $\beta = -1.721$ ,  $SE = 0.708$ ,  $p = .0151$ , 95%CI [-3.109, -3.334e-01]). The interaction term in the givenness and topichood of embedded NP models did not reach significance.

### License

*Language Development Research* (ISSN 2771-7976) is published by TalkBank and the Carnegie Mellon University Library Publishing Service. Copyright © 2026 The Author(s). This work is distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 International license (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits any use, reproduction and distribution of the work for noncommercial purposes without further permission provided the original work is attributed as specified under the terms available via the above link to the Creative Commons website.