Iconicity affects children's comprehension of complex sentences: the role of semantics, clause order, input and individual differences

Laura E. de Ruiter[a] *, Anna L. Theakston[a], Silke Brandt[b], Elena V. M. Lieven[a]

[a] ESRC International Centre for Language and Communicative Development, School of Health Sciences, University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom, Email: {laura.deruiter, anna.theakston, elena.lieven}@manchester.ac.uk

[b] ESRC International Centre for Language and Communicative Development, Linguistics and English Language, Lancaster University, Lancaster, LA1 4YL, United Kingdom, Email: s.brandt@lancaster.ac.uk

* Corresponding author

Acknowledgments

This work was supported by the International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. The authors declare no competing financial interests. The authors would like to thank Moss Hey Primary School in Stockport and all other nursery and primary schools, as well as the children and adults who participated in the study. The authors also thank Yu Li, Michelle Davis, and Felix Engelmann for their help at various stages of the study design, and two anonymous reviewers for helpful comments on an earlier version of this article.

Abstract

Complex sentences involving adverbial clauses appear in children's speech at about three years of age yet children have difficulty comprehending these sentences well into the school years. To date, the reasons for these difficulties are unclear, largely because previous studies have tended to focus on only sub-types of adverbial clauses, or have tested only limited theoretical models. In this paper, we provide the most comprehensive experimental study to date. We tested four-year-olds, five-year-olds and adults on four different adverbial clauses (*before, after, because, if*) to evaluate four different theoretical models (semantic, syntactic, frequency-based and capacity-constrained). 71 children and 10 adults (as controls) completed a forced-choice, picture-selection comprehension test, providing accuracy and response time data. Children also completed a battery of tests to assess their linguistic and general cognitive abilities. We found that children's comprehension was strongly influenced by semantic factors – the iconicity of the event-to-language mappings – and that their response times were influenced by the type of relation expressed by the connective (temporal vs. causal). Neither input frequency (frequency-based account), nor clause order (syntax account) or working memory (capacity-constrained account) provided a good fit to the data. Our findings thus contribute to the development of more sophisticated models of sentence processing. We conclude that such models must also take into account how children's emerging linguistic understanding interacts with developments in other cognitive domains such as their ability to construct mental models and reason flexibly about them.

Keywords: language acquisition; complex syntax; adverbial clauses; iconicity of clause order; input; language processing

1. Introduction

In order to construct a coherent mental representation of the events described in complex sentences, listeners must be able to interpret connectives to establish the semantic relationship (e.g., temporality – *after, when etc.*, causality – *because, since*, concession – *although, even if* etc.) between the main- and the subordinate clause. An additional challenge for listeners is that in English (and other languages, but not in all) the two clauses can occur in two orders. Compare "She had a cup of coffee before she submitted the paper" and "Before she submitted the paper, she had a cup of coffee". In the first sentence, the clause order reflects the order of events in the real world – it is 'iconic'. In the second sentence, the clause order is reversed.

Although complex sentences involving adverbial clauses appear in children's speech at about three years of age (Diessel, 2004), experimental studies found that children have difficulty comprehending these sentences even at the age of six, nine, or even twelve years (e.g., Emerson & Gekoski, 1980; Johnson & Chapman, 1980; Pyykönen, Niemi, & Järvikivi, 2003). They misinterpret the temporal order, or reverse cause and effect in causal sentences. Researchers have suggested different explanations to account for these – often conflicting – findings. But because individual studies have typically looked at only one type of adverbial clause, and used varying methodologies, it is difficult to determine possible differences and commonalities in the precise influences of different factors on children's performance across sentence types. The present study investigates the comprehension of four different sentence types (*after, before, because, if*), to test the predictions of four different theoretical accounts.

We first provide a brief characterisation of the four sentence types under investigation, together with a short discussion of causality, which is central for the understanding of *because*- and *if*-clauses. We then present four different theoretical accounts of complex sentence processing in children that we have identified in the literature: 1) the semantic account, which assumes that iconicity is the main factor; 2) the syntactic account, which assumes that main-subordinate clause orders are easier to process; 3) the frequency-based account, which assumes that forms that are more frequent in the input should be easier to process; 4) the capacity-constrained account, which assumes that individual working memory capacities determine sentence-processing performance. We discuss the details of these four accounts and review the empirical evidence for each of them by summarising previous findings on children's comprehension of sentences containing the connectives *after*, *before*, *because*, and *if*, as well as the few studies done with adult participants.

1.1. COMPLEX SENTENCES

Complex sentences consist of a main and a subordinate clause. While there are other types of complex sentences (e.g., relative clauses, complement clauses), in the context of this article we mean sentences with adverbial clauses.. The adverbial clause is introduced with a connective (subordinating conjunction) that specifies the semantic relationship between the two clauses. In sentences with *before* and *after*, this relationship is purely temporal (indicating priority and posteriority, respectively). Sentences with *because* and *if*, however, can express a range of different meanings. As the present study focusses on one particular type of causality expressed by *because*- and *if*-sentences, we give a short overview of the different types of causality.

According to Sweetser (1990), causality can occur on three different cognitive levels. Compare the utterances in (1-3) below:

1) The cup broke because it fell off the table.
2) She must be a queen, because she is wearing a crown.
3) Can you tell me what time it is, because I have this meeting at one.

In (1), there is a clear causal relation between the two events, and the two events take place in the world independent of the speaker. This type of causality has been called physical or content-level causality. In (2), in contrast, the speaker is using the because-clause as evidence for her (subjective) belief. This type of causality is said to take place on the epistemic level (epistemic causality). Finally, in (3), the because-clause functions as a reason for the speaker's request – it takes place on the level of the speech act (speech act causality). Other scholars have suggested dichotomous distinctions such as objective (content) vs. subjective (epistemic and speech-act) causality (Lois Bloom & Capatides, 1987).

Like *because*-sentences, *if*-sentences can be used to express content-relations, epistemic relations, and speech act relations between clauses. In the content domain, *if*-sentences typically express causal relations via predictions (Dancygier & Sweetser, 2000: 121), as in "If you take this, you'll feel better".

Our study investigates children's comprehension of sentences expressing content-level or physical causality. Note that in this case, there is also a clear temporal element in the semantic relationship between the two events: The cause precedes the effect. However, it is worth pointing out that in conversation, describing causally linked events is not the primary function of *because*- and *if*-sentences. In spoken discourse, *because*-clauses typically provide a reason for a statement made (speech-act causality), rather than a cause for an effect (Diessel & Hetterle, 2011). And *if*-clauses often provide a conceptual framework for the interpretation of the following discourse, not just the main clause within the complex sentence (e.g., Ford & Thompson, 1986). For example, a speaker may say: "If the weather is good tomorrow, we could go for a hike", before providing more details for that proposal. We will return to this distinction between the semantics of *because*- and *if*-clauses and their communicative function at various points in this article.

As noted above, in English, complex sentences can occur in two clause orders: main-subordinate and subordinate-main. (Note that this is true only for adverbial sentences, not for other types of complex sentences.) For each sentence type (*after, before, because, if*) one clause order reflects the order of events in the real world, while the other reverses it. Table 1 illustrates the interaction of connective and clause order yielding (non-) iconicity. For *after*-, *because*-, and *if*-sentences, subordinate-main clause orders are iconic. For *before*-sentences, however, main– subordinate clause orders are iconic.

Connective	Clause order		Iconicity
after	subordinate-main	After he pats the dog, he jumps the gate.	iconic
	main-subordinate	He jumps the gate after he pats the dog.	non-iconic
before	subordinate-main	Before he jumps the gate, he pats the dog.	non-iconic
	main-subordinate	He pats the dog before he jumps the gate.	iconic
because	subordinate-main	Because she puts a hat on, she feels warm.	iconic
	main-subordinate	She feels warm because she puts a hat on.	non-iconic
if	subordinate-main	If she puts a hat on, she feels warm.	iconic
	main-subordinate	She feels warm if she puts a hat on.	non-iconic

Table 1: Interaction of connective type and clause order yielding iconicity.

Iconicity is the central aspect in the semantic account of children's comprehension of complex sentences, which is the first of four different accounts, to which we turn now.

1.2. THEORETICAL ACCOUNTS

1.2.1. Semantic account

Clark (1971) conducted the first experimental study on the acquisition of the temporal connectives *before* and *after*, looking at both production and comprehension in three- to five-year-olds. In the comprehension task, children were asked to act out sentences like "He patted the dog after he jumped the gate" with toys. Not surprisingly, younger children made more errors than older children. In addition, children of all age groups made more errors with those sentences that were non-iconic, and more errors with sentences containing *after* than with sentences containing *before*. These findings led her to suggest that children's comprehension of complex sentences is driven primarily by a semantic principle. Children initially employ an "order-of-mention" strategy: They assume that what they hear first, happens first. In other words, a sentence is being interpreted by assuming a direct mapping (analogy) between the sequence of events in the linguistic form (clause order) and the sequence of events in the real world. As a consequence, children interpret iconic sentences correctly, but misinterpret non-iconic sentences. A correct understanding of both orders emerged in her sample at around age five. It should be pointed out that Clark based her account on an experiment that included only temporal clauses, and did not specify to what extent it should also apply to other complex sentence types. However, it seems reasonable to assume that if children operate with an order-of-mention strategy on the incoming speech stream, they would do so also with causal and conditional sentences, where these describe a causal relationship between two events.

Clark furthermore suggested that *before* and *after* differ in terms of their semantic features. The underlying assumption is that words are made up of a number of semantic features, which can have positive or negative values, such as [±Prior]. In this framework, it is assumed that *after* is more complex than *before* (see E. V. Clark, 1971, for details), which results in an asymmetric acquisition of the two sentence types. Children would start out with wrongly interpreting *after* as *before*.

Subsequent studies that went on to test Clark's hypotheses used a variety of different methods and investigated different age groups (see Table 2), and produced contradictory results. Regarding the comprehension of iconic vs. non-iconic sentences, several studies, including recent ones, have observed better performance with iconic sentences (Blything & Cain, 2016; Blything, Davies, & Cain, 2015; Feagans, 1980; French & Brown, 1977; Stevenson & Pollitt, 1987; Trosborg, 1982, for Danish), although the strength of the evidence is limited for some studies by the fact that they did not manipulate clause order (i.e., order of main- and subordinate clause) (Feagans, 1980), or confounded clause order with plausibility (French & Brown, 1977). Other studies, however, failed to find an advantage for iconic sentences (Amidon, 1976; Gorrell, Crain, & Fodor, 1989; Keller-Cohen, 1987).

Study	**Connective**	**Ages (yrs.)**	**Task(s)**
Amidon, 1976	*after, before, if*	5, 7, 9	Command task (“Before you move the plane…”)
Amidon & Carey, 1972	*after, before*	5-6	Command task (“Before you move the plane…”)
Blything & Cain, 2016	*after, before*	3-7	Forced-choice “what happened last?” (animations)
Blything, Davies, & Cain, 2015	*after, before*	3-7	Forced-choice “what happened first?” (animations)
Carni & French, 1984	*after, before*	3,4	Answering questions after listening to stories
Clark, 1971	*after, before*	3-5	Act-out
Corrigan, 1975	*because*	3-7	sentence-completion, truth-value judgment
Emerson, 1979	*because*	5-8	Forced-choice (picture sequences)
Emerson, 1980	*if*	5-8	Acceptability judgment
Emerson & Gekoski, 1980	*because, if*	3-12	Imitation, forced-choice (picture sequences), recognition, synonymy judgment
Feagans, 1980	*after, before*	3, 5, 7	Act-out
French & Brown, 1977	*after, before*	3-5	Act out
Gorrell, Crain, & Fodor, 1989	*after, before*	3-6	Command task
Johnson, 1975	*after, before*	4-5	Act-out, command task
Johnson & Chapman, 1980	*because*	6, 9, 11	Acceptability judgments, recall
Keller-Cohen, 1987	*after, before*	3-5	Act-out
Kuhn & Phelps, 1976	*because*	5-8	Forced choice (picture sequences)
Stevenson & Pollitt, 1987	*after, before*	3-4	Act-out
Trosborg, 1982	*after, before*	3-7	Act-out, answering questions

Table 2: Overview of previous studies on children's comprehension of complex sentences, indicating the connectives studied (only those relevant for the present study), ages covered (rounded), and tasks used.

Regarding the difference between the two connectives *before* and *after*, previous research has, again, produced divergent results. In line with Clark’s original findings, several studies have found moderate to strong advantages for *before* (Blything & Cain, 2016; Blything et al., 2015; Feagans, 1980; Johnson, 1975), including faster response times in a picture-selection task to sentences containing *before* (Blything & Cain, 2016), while others either did not observe a significant difference between the two (Amidon, 1976; Amidon & Carey, 1972; French & Brown, 1977; Gorrell et al., 1989; Johnson, 1975), or found the opposite, that is, *after* being acquired earlier/being easier than *before* (Carni & French, 1984).

For *because*- and *if*-sentences, the evidence supporting the semantic account is even less clear. This is in part due to methodological issues. On the one hand, many of the studies had relatively high task demands such as requiring meta-linguistic judgments (Corrigan, 1975; Emerson, 1980; Johnson & Chapman, 1980). On the other hand, many used sentences that were constrained by world-knowledge and plausibility (e.g., Kuhn & Phelps, 1976). In order to gauge children's purely linguistic understanding of the meaning of a connective, it is necessary to remove any cues that could guide their interpretation other than the sentence itself. Emerson (1979) addressed this by using so-called reversible sentences, that is, sentences whose reversed meaning is also plausible. She presented children between 5;8 and 10;11 with two different three-frame picture sequences, one corresponding to the order of events in the test sentence, and one showing the opposite order. The children's task was to select which of the two sequences went with the test sentence. Emerson found that children performed better with iconic sentences in which the cause preceded the effect (e.g., "Because he could hear the loud noises and the laughing he went outside"). Only the eight-year-olds were able to make correct selections with non-iconic sentences (e.g., "He went outside because he could hear the loud noises and the laughing"). Emerson and Gekoski (1980) used the same methodology to test the comprehension of *because*- and *if*-sentences in children between 2;8 and 11;11 years, complemented by additional tasks such as asking children to judge the equivalence of meaning in sentences with different connectives (*because/so*, *if/then*) or clause orders. Again, above-chance performance was found only at around eight years, but unlike Emerson's (1979) study, they did not find any effect of iconicity. Amidon (1976) who used a command-task ("If the light comes on, you move the car") similarly found no evidence for an iconicity-preference with *if*-sentences in five-to-nine-year-olds, but she found above-chance performance already in the youngest age-group.

To summarise, there is some, albeit not unequivocal, evidence in support of the semantic account in children: Children seem better at comprehending iconic temporal sentences, and there is some evidence that *before*-sentences may be acquired earlier/be easier to process than *after*-sentences. The role of iconicity for *because*- and *if*-sentences is, however, less clear.

We are aware of only three studies that explicitly studied adult processing of isolated sentences containing *after* and *before*, one study that looked at *because*, and as yet no study using *if*. H. H. Clark and E.V. Clark (1968) gave participants sentences like "After he tooted the horn, he swiped the cabbages" to memorise, together with a noun cue ("the boy"). Participants were then presented with only the noun cues and asked to recall the corresponding sentence. They found that recall was better with iconic sentences. Smith and McMahon (1970) replicated these findings. Münte and colleagues (Münte, Schiltz, & Kutas, 1998) used event-related brain potentials (ERPs) to investigate listeners' processing of sentences with *before* and *after*. Critically, they only compared two types of sentences with each other: iconic *after*-sentences and non-iconic *before*-sentences. They observed that the before-sentences elicited greater negativity, and that the size of the effect was correlated with individual working-memory spans, with individuals with higher spans showing larger negative effects. Münte et al. suggested that this reflects the differential involvement of working memory during the processing of iconic and non-iconic sentences. However, given that clause order and connective type were confounded with iconicity, it is unclear if the observed effect can be attributed to iconicity alone. Finally, in a study on reading comprehension, Irwin (Irwin, 1980) found that college students' answers to multiple choice questions after reading were more accurate if the causal statements (“Because…,) were in iconic form.

In sum, previous adult studies provide some support for the semantic account, but it needs to be pointed out that most of them were reading studies, and that the one study that did use auditory stimuli (Münte et al., 1998) had methodological flaws.

1.2.2. Syntactic account

A competing hypothesis is that the comprehension of complex sentences is mainly affected by syntactic form. Specifically, Diessel (2005) suggested that not only children, but listeners in general, find main-subordinate orders easier to process. For this proposal, he adapted Hawkins’ ‘‘performance theory of order and constituency’’ (Hawkins, 1990, 1992, 1994). In a nutshell, Hawkins assumes that certain syntactic configurations make it easier for the parser to recognise the structure it is currently parsing and to build a hierarchical syntactic representation. In the case of complex sentences, initial connectives like “after” as in 4(a), signal that the structure is a complex sentence. According to Diessel (2005), this requires the parser to keep the subordinate clause in memory until the main clause can be parsed and the complex sentence fully constructed. In 4(b), in contrast, the main clause can be fully processed first. When the subordinate clause is encountered, it can be parsed and attached directly to the representation.

a) [[After he pats the dog]$_{\text{subordinate clause}}$[he jumps the gate]$_{\text{main clause}}$] $_{\text{sentence}}$

b) [[He jumps the gate]$_{\text{main clause}}$[after he pats the dog]$_{\text{subordinate clause}}$]$_{\text{sentence}}$

Main-subordinate orders are thus easier to process, because they have a shorter “recognition domain”: Fewer words must be parsed in order to recognise the syntactic structure of the sentence (see Hawkins, 1992, p. 48 for a formal definition).

Diessel (2005, 2008) acknowledged that in *production*, factors other than syntactic structure play a role in determining the clause order, namely discourse-pragmatic forces, and semantics (iconicity). From a pure processing perspective, however, listeners should find isolated complex sentences easier to process if they occur in main-subordinate order.

To our knowledge there has been no language acquisition study that found support for this hypothesis. Some of the earlier studies cited above, which did not produce corroborative evidence for the semantic account, reported that children appear to understand main clauses better than subordinate clauses (Amidon, 1976; Amidon & Carey, 1972; Gorrell et al., 1989; Johnson, 1975; Stevenson & Pollitt, 1987), but not that main-subordinate orders were comprehended better. While these findings do not support Diessel's hypothesis, they could be taken to indicate that syntax, more specifically, syntactic constituency (main vs. subordinate) plays a role in children's sentence comprehension. It is notable, however, that all the studies that reported a "main clause effect" used a version of the command-task mentioned before (e.g., "Before you move the blue plane, move the red plane"). It was observed that the majority of errors in the children's responses were errors of omission, rather than reversal errors, as observed by Clark and others who used the act-out paradigm. Specifically, children tended to omit the command given in the subordinate clause. Researchers have pointed out that the results may be due to the infelicitous use of a sentence like "Before you move the blue plane, move the red plane" in the experimental set-up. Sentences like these could be "used only when the hearer has established the intent to perform the action mentioned in the subordinate clause" (Gorrell et al., 1989: 625). If this presupposition condition were not met (i.e., if the action in the subordinate clause is not part of the common ground), children would simply ignore this part of the complex sentence. . What appears at first sight to be a syntactic effect is thus probably more likely a pragmatic one.

For adults, Clark and Clark's (1968) study on recall of *before*- and *after*-sentences found – in addition to iconic orders being recalled better than non-iconic ones – that participants performed better with main-subordinate orders. Unlike the iconicity-effect, this facilitative effect of clause order was, however, not replicated by Smith and McMahon (1970).

Overall, the evidence for the syntactic account as put forward by Diessel (2005) is not very strong.

1.2.3. Frequency-based account

Usage-based approaches to language acquisition posit that children's acquisition of grammatical structures is influenced by the frequency of these structures in the children's language input (for an overview, see De Ruiter & Theakston, 2017). Frequency-effects have been observed for a range of syntactic constructions. A frequency-based account would predict that the frequency of order combinations in connective clauses in the input affects children's comprehension of adverbial clauses. Specifically, one would expect that children find those connectives and order combinations which are more frequent easier to understand than those that are less frequent. Both analyses of general language corpora (Diessel, 2001, 2008) and corpora of child-directed speech (De Ruiter, Theakston, Brandt, & Lieven, 2017) have found that

a) *because*- and *if*-sentences are much more frequent than *after*- and *before*-sentences, and
b) there are clear clause order preferences for three of the four sentence types:
    - *if*-sentences occur primarily in subordinate-main order;
    - *before*- and *because*-sentences occur primarily in main-subordinate order.

For *after*-sentences, the picture is less clear. Some studies found that they occur more often in main-subordinate order (Diessel, 2008), others found a preference for subordinate-main orders (De Ruiter et al., 2017.; Diessel, 2005).

If input-frequency influences processing, children (and possibly adults) should find *because*- and *if*-sentences easier to process, and should show facilitative effects for the preferred clause orders of each sentence type, all else being equal. Note that with respect to clause order, the semantic account and the frequency-based account make the same predictions for *before*, and *if*-clauses, because for these sentences, the iconic clause order is also (and probably not accidentally, e.g., Diessel, 2005) the most frequent one. Different predictions emerge for *because*-sentences, however: While the semantic account predicts that sentences beginning with *because* are easier to process (subordinate-main), the frequency-based account would predict that sentences in which the *because*-clause follows the main clause are easier to process/acquired earlier.

However, frequency effects can occur on different levels of abstraction. Children and adults are also sensitive to discourse-based and semantic features of lexical items that are most frequently used in specific constructions. For example, Kidd and colleagues (Brandt, Kidd, Lieven, & Tomasello, 2009; Kidd, Brandt, Lieven, & Tomasello, 2007) found that children most often hear object relative clauses with inanimate head nouns and pronominal subjects, and they also understand these complex sentence types best when the sentences are formed according to these constraints. A prototypical feature of complex sentences is that they contain transitive verbs (De Ruiter et al., 2017.). One would thus expect that complex sentences with transitive verbs pose fewer difficulties for children than sentences with intransitive verbs.

There have – to our knowledge – not been any investigations of the links between input frequencies of complex sentence forms with adverbial clauses and children's comprehension of these sentences. However, with the corpus findings regarding the different frequencies of connectives and clause orders in mind (see above), we can evaluate the results of previous studies. The only study that covered three of the four connectives (*after*, *before*, and *if*) found that five-to-nine-year-old children showed overall lower error-rates with *if*-sentences than with *after*- and *before*-sentences (Amidon, 1976), in support of a frequency-based account. Moreover, to the extent that children have shown a tendency to perform better with iconic sentences (see semantic account above), these sentences also reflect the more frequent clause orders for *before* and *if*-sentences, (and possibly *after*-sentences) in spoken English. But the evidence for *because*-sentences, which occur more often in (non-iconic) main-subordinate orders, is rather sketchy. On the other hand, the approximate ages at which children have been reported to perform above-chance in their comprehension of complex sentences in the various studies indicate that *because*- and *if*-sentences may show a more protracted development than *after*- and *before*-sentences. This would seem to contrast with what would be predicted on the basis of a pure form-frequency-based account.

1.2.4. Memory capacity-constrained account

Theories of capacity constraints in memory (e.g., Just & Carpenter, 1992) assume that short-term memory[1] plays a central role in sentence processing, and, crucially, that there are individual differences in the resources that a listener (or reader) has at their disposal. As a consequence, individuals with lower memory capacity will find it more difficult to keep more information in active storage during parsing.

Note that the capacity-constrained account is not compatible with the semantic account, because children's use of the iconicity principle (and the semantic features account) is not assumed to be linked to memory in any way. The capacity-constrained account is, however, in theory compatible with both the syntactic and the frequency-based account. The syntactic account makes explicit predictions about the processing difficulty associated with the two clause orders. It is possible that difficulties with subordinate-main orders are exacerbated by low short-term memory capabilities. The frequency-based account does not say anything

[1] While Just & Carpenter use the term "working memory", we prefer to describe the capacity involved as "short-term memory", because the task doesn't involve manipulation of the stored information. But the two terms are often used interchangeably, and researchers have difficulties separating the two constructs (see Aben, Stapert, & Blokland, 2012 for a discussion).

about the influence of memory, but there is no a-priori reason why frequency-effects could not be modulated by working memory. For the syntactic and the frequency-based account, then, the capacity-constrained account provides an additional hypothesis, rather than an alternative: Children with better working memory should perform better in complex sentence comprehension tasks than children with lower working memory capabilities.

Blything and Cain (2016), who investigated three- to seven-year-old children's comprehension of sentences with *before* and *after*, found some support for the capacity-constrained account. Performance in terms of accuracy and speed (response time) was predicted better by children's scores on a memory task (digit span) than by age or vocabulary (Blything & Cain, 2016). To our knowledge there have been no studies that examined the link between memory and comprehension of *because*- and *if*-sentences. Studies that investigated the role of working memory in the processing of other types of complex sentences (e.g., passives, relative-clauses) have found that memory significantly predicted sentence comprehension over and above the influence of age (Magimairaj & Montgomery, 2012; e.g., Montgomery, Magimairaj, & O'Malley, 2008).

For adults, Münte and colleagues (1998) found that participants with higher working memory spans showed a more pronounced difference between *before*- and *after*-sentences in terms of ERP negativity. They took this to indicate that these participants were probably better comprehenders, although the study did not directly measure comprehension.

Taken together, there is some evidence that individual memory capacities influence complex sentence processing in general, but up to this point there is only limited support for this hypothesis for adverbial clause processing specifically.

To sum up: There are four different theoretical accounts for the comprehension of complex sentences: the semantic account, the syntactic account, the frequency-based account, and the capacity-constrained account. More than four decades of research have produced some support for each of the four accounts, but because researchers have typically focussed on certain types of sentences, and used a plethora of different methods, it is difficult to decide between them. Our study evaluates and compares the predictive adequacy of these different accounts. We also consider how they may interact in the Discussion.

1.3. The present study

Our study tests the predictions made by different theoretical accounts across four different sentence types (*after*, *before*, *because*, *if*) by using the same methodology (forced-choice, picture-sequence selection) for all types and testing the same children (within-subjects design), as well as including measures of short-term memory. Because it is unclear what the role of individual differences in general language ability and executive function may be in complex sentence comprehension, and in order to control for potential confounding factors, we furthermore collected measures of general language ability and executive function (inhibition). We also tested children's understanding of the temporal priority principle (causality). If the children in our sample generally understand (event) causality, then a failure to comprehend the causal sentences must be due to a lack of linguistic rather than conceptual knowledge. In addition, we tested an adult control group to provide a baseline/.

2. Materials and Methods

2.1. Participants

Seventy-one children and ten adults participated. The children were recruited through nurseries and primary schools in the North-West of England. Prior informed consent was obtained from caregivers/parents. All children were monolingual, native speakers of English without any known history of speech or language problems or developmental delays. Of the 71 child participants, 37 were between 3;6 to 4;5 years old (M = 47 months, SD = 3.8, 20 girls), and 34 were between 4;6 and 5;5 years old (M = 60 months, SD = 3.1, 25 girls). We will refer to the first group as the four-year-olds, and the second group as the five-year-olds.

Eight additional children were tested, but their data had to be excluded because they turned out to be bilingual (three participants), too old (two participants), too young (one participant), or because they did not understand the task (two participants). One child refused to do the second session, while the second session with another child had to be aborted shortly before completion due to concentration problems, resulting in loss of two responses. A technical failure caused the loss of three responses with another participant. Half of the data set of one child was lost due to experimenter error. The adult participants (N = 10, M = 33 years, seven women) were students or staff members at a university in the North-West of England, and native speakers of English.

2.2. *Materials and Procedure*

The children were tested in a quiet area in their nurseries and primary schools. In addition to the sentence comprehension test, children completed five tasks on general language ability, short-term memory, executive control, and understanding of causality (all detailed below), spread over two sessions on two days. Each session lasted between 25 and 40 minutes. Children completed half of all items of the sentence comprehension task in session one, and the other half in session two. The language ability tasks and the executive control tasks were administered in session one. The memory test and the causality test were administered in session two. In both sessions, children always first completed the sentence comprehension task before doing the other tasks. The allocation of trials across sessions and the experimental lists are described in *Experimental lists* below. Adult participants did only the sentence comprehension task and completed all items in one session, with a short break between the two blocks.

1.1.1. Sentence Comprehension

Participants' comprehension of complex sentences was tested using a forced-choice picture-sequence selection task on a touch-screen. The task was to select out of two picture sequences the one that matched an aurally presented sentence. This allowed us to collect both response accuracy and reaction time measures.

1.1.1.1. Design

The experiment had four factors: one between-subjects factor (AgeGroup), and three within-subjects factors (Type, ClauseOrder, VerbType), each with the following levels:

- AgeGroup: 4 years, 5 years
- Type: after, before, because, if
- ClauseOrder: main-subordinate, subordinate-main
- VerbType: transitive, intransitive

Table 3 shows examples of stimuli in the different conditions.

Connective	after		before		because		if	
Clause order	main-sub	sub-main	main-sub	sub-main	main-sub	sub-main	main-sub	sub-main
Transitive verbs	*She hoovers the house after she paints the old fence.*	*After she paints the old fence, she hoovers the house.*	*He plays his big drum, before he reads his new book.*	*Before he reads his new book, he plays his big drum.*	*He opens the door, because he sees the snowman.*	*Because he sees the snowman, he opens the door.*	*She hears the doorbell, if she presses the button.*	*If she presses the button, she hears the doorbell.*
Intransitive verbs	*He drives away fast after he shouts out loudly.*	*After he shouts out loudly, he drives away fast.*	*She hops up and down before she crawls on the floor.*	*Before she crawls on the floor, she hops up and down.*	*She slips to the ground, because she looks at the sky.*	*Because she looks at the sky, she slips to the ground.*	*He falls in the field, if he sneezes lots of times.*	*If he sneezes lots of times, he falls in the field.*

Table 3: Conditions of the experiment, 4 connectives x 2 clause orders (main = main clause, sub = subordinate clause) x 2 verb types (transitive, intransitive)

For the adult group, there were three within-subjects factors (Type, ClauseOrder, VerbType). There were three items per condition, 48 items overall.

1.1.1.2. Audio stimuli

24 complex sentences were constructed, each containing a main and subordinate clause representing two actions performed by a single actor (a boy in half of the sentences, and a girl in the other). There were six sentences per connective *after, before, because*, and *if*. The *because*- and *if*-sentences always expressed a physical causal relationship between the two events (i.e., not epistemic or speech act relations). The stimuli clearly emphasised the causal interpretation of these sentences (there was always only one person in each scene,

making the use of speech act-causality implausible). Within these six sentences, half (three) contained only intransitive verbs, the other half contained only transitive verbs. The objects of the transitive verbs were always inanimate objects. Each sentence occurred in both clause orders (main-subordinate and subordinate-main), resulting in 48 sentences overall. The subject of the sentence was always expressed as a pronoun (i.e., *he* or *she*), and all verbs were in present tense. All sentences were between 11 and 13 syllables long. (All experimental sentences can be found in Table A 1 in Appendix A.)

The sentences were spoken by a female native speaker of British English, and recorded in a quiet room using a digital voice recorder. The stimuli were processed using the software Praat (Boersma & Weenink, 2016), version 6.0.13. Each sentence was first cut into two clauses, and then spliced together again with a pause of 250ms. The overall intensity of all stimuli was set to 60dB.

1.1.1.3. Visual stimuli

For each audio stimulus (complex sentence), two picture sequences were created (for an example, see Table 4), showing the two actions expressed by the sentence in both orders (in left-to-right orientation, which is the convention in English picture books). For the sentences containing *before* and *after*, the second picture sequence was the reversal of the pictures of the first picture sequence. This was not possible for the sentences containing *because* and *if*, since the semantics of these sentences requires there be some change of state involved. For example in the sequence matching the sentence “Because he opens the door he sees the snowman”, the actor first opens the front door and then finds a snowman outside his house. The other sequence has to offer a plausible scenario for the opposite order of events (i.e., first seeing, then opening) in order to be an acceptable distractor. In this case, the actor was depicted as looking out of the window and seeing a snowman, and then opening the door (to have a better look at the snowman). The stimuli were created using the software Anime Pro (version 9.1).

1.1.1.4. Presentation

The stimuli were presented using the software E-Prime (version 1.2) on a laptop with a 14-inch resistive touch-screen. The sound was presented via loudspeakers.

1.1.1.5. Procedure

Children

The children sat at a table in front of the laptop. In front of the laptop there were two pieces of red cardboard in hand shape fixed to the table. The children were asked to keep their hands on these markers throughout the experiment when they were not selecting a sequence. The children were told that they were going to play a game, in which a lady was telling them stories about two characters, Sue and Tom, and about some animals, and that they had to select from two picture stories the one that matched the sequence that they had heard. The children were instructed to listen carefully and touch the matching sequence after they hear a beep.

Before the start of the actual experiment, there was a warm-up phase to familiarise the children with the task and the left-to-right reading of the picture sequences. In the warm-up, the second presentation of the sentence (see below for details of the set-up) was not automatic, but manually controlled by the experimenter, which allowed the experimenter to explain the layout of the screen before playing the sentence again (e.g., "Here we see that Tom is doing two things in this story. First he is watering his plants. And then he switches the light on", while pointing to the appropriate picture). The first two warm-up trials were like the filler trials (i.e., simple sentences with only two pictures; see below). The other warm-up trials were like the experimental trials, except that the sentences were of the structure "First, …, then…". If a child did not choose the correct picture in any of the warm-up trials, feedback was given and the trial was repeated up to two times. If the child still made the wrong selection, the experimenter proceeded to the experimental trials, but noted that the child had failed to complete the warm-up successfully.

The structure of the experimental trials is shown in Table 4. Before each trial, there was a picture of the character that the next “sequence” was about (i.e., a picture of Sue or Tom). The experimenter would say something like “Ah, here’s another story about Sue. Let’s see what she’s doing!” to focus the child’s attention on the next trial. When the experimenter was sure that the child was paying attention, she started the next trial. The child would first hear the instruction “Look and listen carefully! Touch the matching story after the beep!”[2], while seeing a blank screen. Then the sentence was played, with the screen still blank. Directly after the presentation of the sentence, the two picture sequences were displayed on the screen. After a pause of 1000 ms, the sentence was repeated, followed immediately by a beep. Once the child had selected a sequence, the screen showed a blue circle to indicate that the trial had been successfully completed. Response time was measured from the offset of the beep. If the child was distracted during a trial, the experimenter repeated the trial.

[2] One reviewer remarked that, while it is rather unlikely, using the word “after” in the instructions might have positively impacted the children’s performance. The results suggest that this was not the case, as the children’s performance on *after* was worse than with *before*.

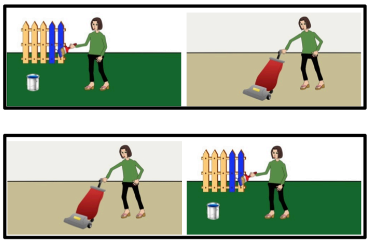
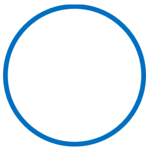
Visual presentation	Auditory presentation
blank screen	"Look and listen carefully! Touch the matching story after the beep!"
	"After she paints the old fence, she hoovers the house."
	*1000 ms pause*
	"After she paints the old fence, she hoovers the house."
	beep
	

Table 4: Structure of the experimental trials.

After every three trials there was a filler trial to give children a small break with relatively easier items. The structure of the filler trials was the same as that of the experimental trials, the difference being that children were presented with a simple sentence (e.g., "Lion is drying his hair.") and only two pictures to select from (e.g., a lion drying his hair and a lion buttoning his coat).

The entire experiment took between 15 and 20 minutes.

Adults

The adult participants were tested in a quiet room, using the same set-up as with the children. Instead of using the hand-shaped markers adults were simply instructed to keep their hands in front of the laptop unless they were selecting a picture sequence. Participants were instructed to listen to the sentence and select the matching sequence after the beep. The warm-up was the same as with the child participants, but no elaborate explanations were provided. After the participants had successfully completed the warm-up, they went through half of the trials, followed by a short break, and then completed the other half of the trials. Overall the experiment took about 10-15 minutes.

1.1.1.6. Experimental lists

Four different experimental lists were constructed. Each list consisted of two sessions. Each sentence (N=24) occurred once in each session (recall that each sentence occurred in two clause orders), with half of the sentences in each session being in main-subordinate clause order and the other half in subordinate-main clause order. There were three items in each condition. List 2 was created by swapping session 1 and session 2 of List 1. Lists 3 and 4 were the same as Lists 1 and 2, with the difference that all *after*-sentences were turned into *before*-sentences and vice versa, and all *if*-sentences were changed into *because*-sentences and vice versa (see Table A 1 in Appendix A).

The order of the trials within each session was pseudo-randomised. There was a maximum of two consecutive trials in the same condition. The position of the correct picture sequence in session 1 was counterbalanced, so that in half of the trials the correct picture sequence was at the top and in the other half of the trials at the bottom. In addition, the position of the correct picture sequence across sessions was counterbalanced, so that for any given scene, when the correct picture was at the top in session 1, it was at the bottom in session 2, and vice versa.

Participants were randomly assigned to one of the four experimental lists.

1.1.2. Language Ability

Measures for children’s receptive language ability were collected using two sub-tests of the Clinical Evaluation of Language Fundamentals®-Preschool-2 (CELF-Preschool-2 Wiig, Secord, & Semel, 2004): “Linguistic Concepts” and “Sentence Structure”. The sub-test “Linguistic Concepts” requires the child to follow directions of increasing length and complexity (e.g., “Point to either of the monkeys and all of the tigers.”). The sub-test “Sentence Structure” is a forced-choice picture selection task that tests the child’s comprehension of sentences of increasing length and complexity (e.g., “The man who sits under the tree is wearing a hat.”). Each sub-test lasted approximately 5 minutes.

1.1.3. Executive Control

Children’s executive control was tested using two tasks: the “Day/Night task” (Gerstadt, Hong, & Diamond, 1994), and the dimensional change card sort (DCCS) task (Zelazo, 2006). In the Day/Night task, children are instructed to say “day” when they are shown a card with a picture of a moon on it, and to say “night” when shown a card with a picture of a sun on it. The task taps into children’s ability to inhibit the intuitive response (e.g., to say “night” when they see a picture of a moon). In the DCCS task, children are required to sort a series of bivalent test cards, first (pre-switch phase) according to one dimension (colour), and then (post-switch phase) according to the other (shape). The task taps into children’s flexibility to switch their attention to a different dimension. Both tasks together took about 5 minutes (16 trials in the Day/Night task, 12 trials in the DCCS task).

1.1.4. Memory

Phonological and verbal short-term memory was tested using three tasks, taken from the Early Repetition Battery® (Seeff-Gabriel, Chiat, & Roy, 2008): word repetition and non-word repetition (which are combined into the “Preschool Repetition Test”, PSRep), and “Sentence Imitation Test” (SIT). All three tasks together took between 5-10 minutes.

1.1.5. Causality

Children's understanding of the temporal priority principle (i.e., the principle that causes must precede their effects) was tested using a modified version of the set-up used by Rankin and McCormack (2013). Children have to decide which one of two events (A, B) causes an effect (E). In the task, children observe one event (A), an effect (E), and then another event (B). The events A and B are marbles rolling down runways, and the effect E is the ringing of a bell. There were four experimental trials. The task took about 5 minutes.

2.3. *Predictions and Analyses*

Based on the four accounts outlined in the introduction, we list a number of different hypotheses regarding children's performance accuracy in the sentence comprehension task:

1. Iconic clause orders are comprehended better/acquired earlier than non-iconic clause orders. (semantic account)
2. *Before*-sentences are comprehended better/acquired earlier than *after*-sentences. (semantic account)
3. Main-subordinate orders are comprehended better/acquired earlier than subordinate-main orders. (syntactic account)
4. *Because*- and *if*-sentences are comprehended better/acquired earlier than *after*- and *before*-sentences. (frequency-based account)
5. Frequent, connective-clause order combinations are comprehended better/acquired earlier than infrequent ones. (frequency-based account)
6. Sentences with transitive verbs are comprehended better/acquired earlier than sentences with intransitive verbs. (frequency-based account)
7. Memory should make an independent contribution to performance, in that children with higher memory scores perform better than children with lower memory scores. (capacity-constrained account)

The accounts do not make explicit predictions about the speed of processing (response times), but it seems reasonable to assume that those structures that are easier to comprehend would also be processed faster.

3. Results

A total of 3907 responses were recorded. After screening of the data for deviations, the data of one child participant was removed, because he had consistently touched the top right-hand corner of the touchscreen, and also confirmed this when asked about it after the experiment. As a result, 48 responses (1% of the data) were excluded.

3.1. Analysis Strategy

We first present the results for the sentence comprehension task (accuracy and response times). We then present the results (raw scores and standardised scores, where applicable) for the other tasks, and test if the individual difference scores in those tasks explain performance over and above the effects of our experimental manipulations.

For accuracy and response times (RTs), a series of (generalised) linear mixed effect models (GLMMs; Baayen, Davidson, & Bates, 2008) was fitted to the data using R (R Core Team, 2016), version 3.3.1. We used glmer for the binomial accuracy dependent variable, and lmer for the continuous response times (RT) dependent variable, both from the R package lme4 (Bates, Maechler, Bolker, & Walker, 2015). We used the R packages lmerTest (Kuznetsova, Brockhoff, & Bojesen Christensen, 2016) and pbkrtest (Halekoh & Højsgaard, 2014) for the calculation of p-values for lmer models. (G)LMMs allow incorporating both fixed effects (experimental manipulations) and random effects (variation specific to individual participants and individual items). Following Bates et al.'s (2015) recommendations, we added fixed and random effects incrementally to a minimal model, and tested if the inclusion of an additional term was justified using the likelihood ratio test for model comparisons (Pinheiro & Bates, 2000), and pruned non-significant effects, unless they were part of a significant interaction.

All final models contained random intercepts for participants and items. In addition, we ran t-tests to test if performance for the subgroups was above chance. For all other tasks (with the exception of the causality task) we ran simple correlations between (centred) test scores and mean accuracy and RT, respectively.

In addition, we performed Bayesian analyses. The reason for this is that conventional significance tests are designed to reject the null hypothesis. However, if the null hypothesis is true, p-values do not converge to any limit value, and all p-values are all equally likely (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Non-significant results therefore do not allow for inference of the truth of the null hypothesis (see e.g., Dienes, 2014). Bayesian analyses, in contrast, provide information about the strength of statistical evidence in favour of either the alternative hypothesis or the null hypothesis. Bayes factors provide the relative probability of the data under the two hypotheses. For example, a Bayes factor of 2 means that the data are two times more likely under the alternative hypothesis (H_A) than they are under the null hypothesis. Similarly, two statistical models can be compared directly with each other, and the strength of the evidence for one model (that includes a given main effect or interaction) over the other (that does not contain this effect or interaction) can be determined. An overview of a common textual interpretation of Bayes factor values is presented in Table 5.

Bayes factor	Interpretation
> 100	Decisive evidence for H_A
30 – 100	Very strong evidence for H_A
10 – 30	Strong evidence for H_A
3 – 10	Substantial evidence for H_A
1 – 3	Anecdotal evidence for H_A
1	No evidence
1/3 – 1	Anecdotal evidence for H_0
1/10 – 1/3	Substantial evidence for H_0
1/30 – 1/10	Strong evidence for H_0
1/100 – 1/30	Very strong evidence for H_0
< 1/100	Decisive evidence for H_0

Table 5: Evidence categories for Bayes factor, adapted from Jeffreys (1961), cited in Wetzels et al. (2011). H_A = alternative hypothesis, H_0 = null hypothesis.

We used Bayesian linear regression from the BayesFactor package (Morey, Rouder, & Jamil, 2015). This type of analysis allows comparing a number of different models and determining the model that is most likely given the data (that is, the model with the highest Bayes factor), and the incorporation of random factors (participant, item). In line with recommendations by Morey and Rouder (2011) we used a Cauchy prior with scale parameter 1/√2 for the standardized effect size. Cauchy priors are relatively wide and symmetric around zero, which means that the data quickly overwhelms the prior (Morey & Wagenmakers, 2014: 123). In addition, we used Bayesian t-tests (from the BayesFactor package) and Bayesian correlations from the BayesMed package (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2014) to complement the traditional analysis outlined above.

3.2. *SENTENCE COMPREHENSION TASK*

3.2.1. ACCURACY

Summary data, together with the adult comparison data, are shown in Figure 1.

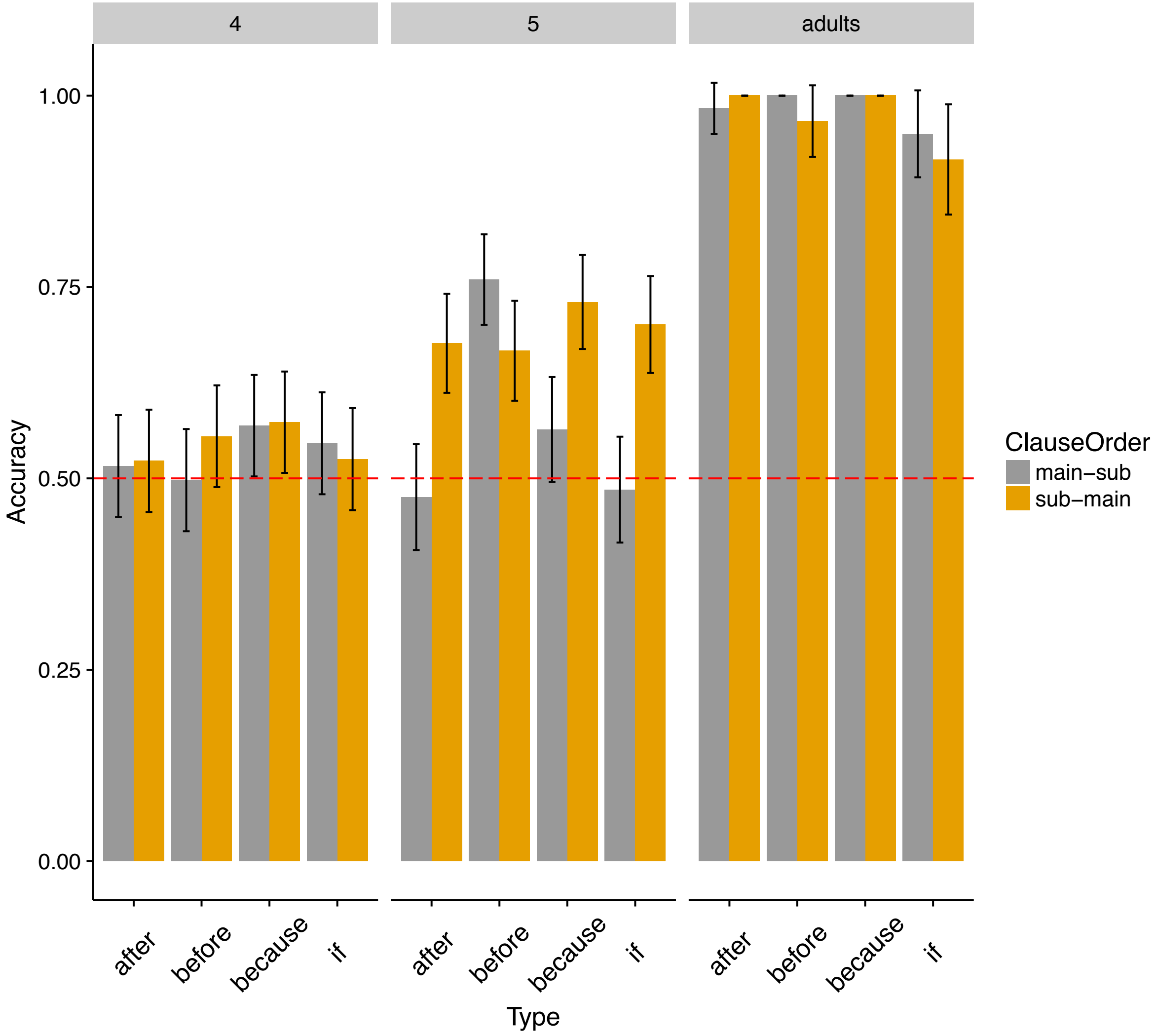


Figure 1: Four-year-olds', five-year-olds' and adults' mean proportion of correct responses for *after*-, *before*, *because*- and *if*-clauses in subordinate-main and main-subordinate clause order. The dashed red line indicates chance level. Error bars indicate standard errors.

The mean accuracy in the four-year-old group was 58.3%. The five-year-olds' mean accuracy was higher, at 63.2%. Adults responded correctly in 97.7% of all trials. The summary of the final (traditional) mixed-effects model is shown in Table 6. It shows that there were no significant main effects of AgeGroup, Type, or ClauseOrder, but there were significant interactions of AgeGroup and Type, AgeGroup and ClauseOrder, as well as a three-way interaction of AgeGroup, Type, and ClauseOrder. VerbType was not a significant factor.

Fixed effects	**Estimate**	**Std. Error**	**z value**	**Pr(>lzl)**
(Intercept)	0.07	0.16	0.43	.67
AgeGroup5	-0.17	0.22	-0.79	.43
Typebefore	-0.07	0.19	-0.39	.70
Typebecause	0.22	0.21	1.03	.31
Typeif	0.13	0.21	0.61	.54
ClauseOrdersub-main	0.03	0.19	0.15	.88
AgeGroup5:Typebefore	**1.38**	**0.29**	**4.71**	**< .0001**
AgeGroup5:Typebecause	0.16	0.28	0.57	.57
AgeGroup5:Typeif	-0.09	0.28	-0.33	.74
AgeGroup5:ClauseOrdersub-main	**0.85**	**0.29**	**2.97**	**< .01**
AgeGroup4:Typebefore:ClauseOrdersub-main	0.21	0.27	0.76	.45
AgeGroup5:Typebefore:ClauseOrdersub-main	**-1.35**	**0.31**	**-4.39**	**< .0001**
AgeGroup4:Typebecause:ClauseOrdersub-main	-0.01	0.28	-0.04	.97
AgeGroup5:Typebecause:ClauseOrdersub-main	-0.10	0.30	-0.32	.75
AgeGroup4:Typeif:ClauseOrdersub-main	-0.12	0.28	-0.42	.67
AgeGroup5:Typeif:ClauseOrdersub-main	0.09	0.30	0.29	.77

Table 6: Summary of Generalized Linear Mixed Effects Model for the log odds for accuracy responses: effects and Interactions of AgeGroup, Type, and ClauseOrder. The reference levels are for AgeGroup: 4years, for Type: after, and for ClauseOrder: main-subordinate. Significant effects are highlighted in bold font. Note that because Type has four levels, this table shows the results only for the model with "after" as the reference level. The summaries of the models with the other three connectives as reference level showing the same effects can be found in Appendix B.

The significant interactions can be interpreted as follows. The five-year-olds performed significantly better than the four-year-olds with *before*-sentences (71.3% vs. 61.7%), and also with sentences in subordinate-main orders overall (69.4% vs. 61.6%). However, for *before*-sentences, the five-year-olds' performance with sentences in subordinate-main order was significantly worse than in main-subordinate order (66.7% vs. 76%). This means that the five-year-olds were generally better with sentences in iconic clause order (subordinate-main for *after, because, if,* and main-subordinate for *before*). Adults performed at ceiling. (Note, however, that there were a few errors with *if*- and *because*-sentences, which were due to one particular item. We return to this in the discussion.)

The results of the GLMM were corroborated by the Bayesian analysis. The model that was most likely, given the data, included the same main effects and interactions (Bayes factor: > 60million – "decisive evidence" –, compared to only the intercept). In fact, the model that included the three-way-interaction was 72 times more likely ("very strong evidence") than the model that did not include this three-way-interaction.

While the five-year-olds' performance with *before*-sentences in general and with all other types in subordinate-main order was clearly above chance, it is possible that the four-year-olds overall, and the five-year-olds in the main-subordinate conditions of the other connectives (*after, because, if*) were at chance level – which would explain the absence of a main effect of AgeGroup. We tested each age group's performance in the eight conditions using one-tailed t-tests and Bayesian t-tests. As statistical significance in null hypothesis testing depends on the number of intended analyses, it is necessary to correct for multiple comparisons. Using Bonferroni-correction, adjusting for 18 comparisons (one for each condition, plus two overall) yielded a significance level of 0.05/18 = 0.0028. A correction for multiple comparisons is not necessary for Bayesian t-tests (Dienes, 2011). The results are presented in Table 7 (four-year-olds) and Table 8 (five-year-olds).

	t-test (p)		**Bayesian t-tests (BF)**	
	main-sub	*sub-main*	*main-sub*	*sub-main*
after	.343	.271	0.18$^{\lozenge}$	0.2$^{\lozenge}$
before	.554	.060	0.16$^{\lozenge}$	0.6
because	.025	.018	1.24	1.64
if	.099	.250	0.4	0.21$^{\lozenge}$
overall	< .001*		9.31$^{\lozenge}$	

Table 7: P-values and Bayes factors for (one-tailed) t-tests testing if performance is above chance in the four-year-old age group. Asterisks indicate statistical significance after Bonferroni correction, diamonds indicate at least substantial evidence (for the H_0, if below 1/3, for the H_A if above 3).

	t-test (p)		**Bayesian t-tests (BF)**	
	main-sub	*sub-main*	*main-sub*	*sub-main*
after	.78	< .0001*	0.21$^{\lozenge}$	49484$^{\lozenge}$
before	< .0001*	< .0001*	254384130705$^{\lozenge}$	12382$^{\lozenge}$
because	.04	< .0001*	0.86	497025782$^{\lozenge}$
if	.69	< .0001*	0.18$^{\lozenge}$	2318159$^{\lozenge}$
overall	< .0001*		6.54	

Table 8: P-values and Bayes factors for (one-tailed) t-tests testing if performance is above chance in the five-year-old age group. Asterisks indicate statistical significance after Bonferroni correction, diamonds indicate at least substantial evidence (for the H_0, if below 1/3, for the H_A if above 3).

The t-tests show that the four-year-olds' performance overall was above chance, but this emerges only when all conditions are combined – none of the individual sentence types were above chance after controlling for multiple comparisons. While the p-values are not statistically significant after correcting for multiple comparisons, the Bayes factors provide more information: They show that there is "anecdotal evidence" for above-chance performance with *because*-sentences in the four-year-olds, which is likely to be the reason for their above-chance performance overall. In addition, the Bayes factors show that there is "substantial evidence" for an at-chance performance of the four-year-olds in all *after*-sentences, in *before*-sentences in main-sub order, and in *if*-sentences in sub-main order, and "anecdotal evidence" for at-chance performance in *before*-sentences in sub-main order, and *if*-sentences in main-sub order. In addition, there is evidence that the five-year-olds' performance in main-sub ordered sentences was at chance for *after*-, *because*- and *if*-sentences.

In summary, four-year-olds showed only a very fragile understanding of complex sentences on this task. Five-year-olds showed a better understanding of sentences that were in iconic clause order, and for *before*-sentences overall.

3.2.2. Response times

For the analyses of RTs, only correct responses were analysed (N=2441). After inspection of the data, we removed outliers using the following criteria: For children, we excluded all responses that were shorter than 300ms and longer than 20000ms (99 responses, 5.9% of the data), as it is unlikely that shorter or longer RTs reflect processing of the target stimuli. For adults, we excluded all responses that were shorter than 150ms and longer than 6000ms (17 responses, 3.6% of the data). Overall, 68% of the data from the full data set were included (50% of the 4-year-olds' data, 59% of the 5-year-olds' data, and 94% of the adult data).

The RT data of all age groups are visualised Figure 2.

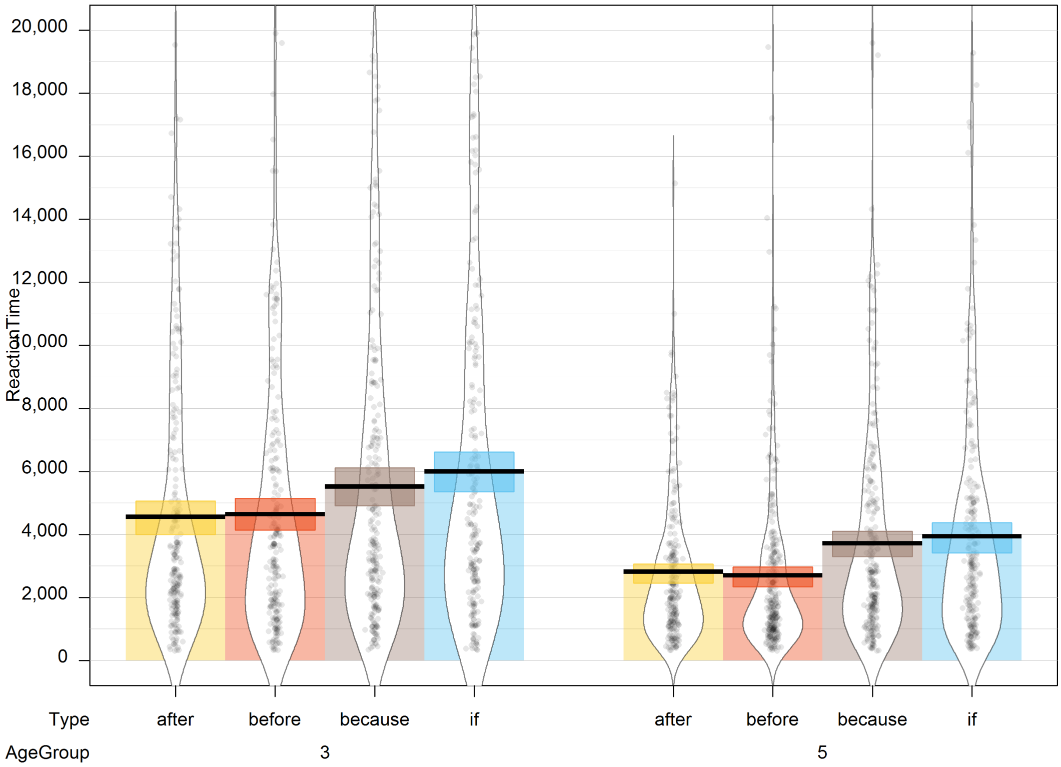


Figure 2: Response times (in milliseconds) of the four-year-olds, the-five-year-olds and the adults for the four different sentence types *after*, *before*, *because*, and *if*. Individual dots represent individual responses (raw data). Bars indicate means, beans (the oval shapes around the dots) indicate smoothed density, and bands (dark-coloured lines at the top of the bars) indicate the 95% Bayesian Highest Density Interval (HDI). The pirate plot was produced using the R package "yarrr" (Phillips, 2016).

The four-year-olds' mean response time was 5177ms, the five-year-olds' was 3278ms, and the adults' 1038ms.

The summary of the final model for the child groups is shown in Table 9. In addition to random intercepts for participants and items the model also contained by-participant slopes for Type. ClauseOrder, and VerbType were not significant factors, but AgeGroup and Type were. There were no significant interactions.

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	4578.29	336.84	105.48	13.471	< .0001
AgeGroup5	**-1750.53**	**417.22**	**99.88**	**-4.095**	**< .0001**
Typebefore	89.33	219.39	74.01	0.405	.69
Typebecause	**1047.36**	**327.32**	**45.44**	**3.193**	**< .01**
Typeif	**1227.54**	**365.56**	**61.14**	**3.353**	**< .01**

Table 9: Summary of Linear Mixed Effects Model for response times for the child groups: effects of AgeGroup and Type. The reference levels are for AgeGroup: 4years, for Type: after. Significant effects are highlighted in bold font. Note that because Type has four levels, this table shows the results only for the model with "after" as the reference level. The summaries of the models with the other three connectives as reference level showing the same effects can be found in Appendix C.

The model was corroborated by the Bayesian analysis: The model under which the data are most likely was the one that contained only AgeGroup and Type as factors (Bayes factor for this model: 5.7, "substantial evidence"). This model was about 19 times more likely than a model that also included ClauseOrder. This provides strong evidence that clause order was not a factor that affected children's response times.

Looking at the effects in the model in Table 9, it can be seen that the five-year-olds responded significantly faster than the four-year-olds. Furthermore, responses to *because*- and *if*-sentences were significantly slower than responses to *after*- and *before*-sentences.

The summary for the model for the adult control group is presented in Table 10. The only significant factor was Type. Adults responded to *before*-sentences significantly faster than to any other sentence-type. However, the Bayesian analysis indicated that the data is about four times more likely under a model with only Participant and Item as random factors than under a model that also contains Type as factor.

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	1083.93	191.2	11.5	5.669	< .001
Typebefore	**-206.43**	**103.46**	**448.1**	**-1.995**	**< .05**
Typebecause	0.99	106.65	54.3	0.009	.99
Typeif	65.43	108.56	55.1	0.603	.55

Table 10: Summary of Linear Mixed Effects Model for response times for the adult group: effect of Type. The reference level is "after". Significant effects are highlighted in bold font. Note that because Type has four levels, this table shows the results only for the model with "after" as the reference level. The summaries of the models with the other three connectives as reference level showing the same effects can be found in Appendix C.

In summary, while neither VerbType nor ClauseOrder had an effect on participants' reaction times, Type had: Children had significantly slower responses with *because*- and *if*-sentences. For adults, it may be the case that *before*-sentences are responded to more quickly, but the results of the two analyses (traditional and Bayesian) are ambiguous.

3.3. INTERIM DISCUSSION

In the introduction, we presented four different theoretical accounts that have been put forward to explain and predict the processing of complex sentences. The semantic account predicts that children will perform better with iconic sentences, and that *before*-sentences will be acquired earlier. The syntactic account predicts that sentences in main-subordinate orders are easier to process. The frequency-based account predicts that *because*- and *if*-clauses should be acquired earlier/more easily processed, and that for a given connective, performance should be better with the more frequently occurring clause order. In addition, sentences with transitive verbs should be easier than sentences with intransitive verbs. Finally, the capacity-constrained account predicts that individuals with better short-term memory skills should perform better generally.

In terms of accuracy, the results showed that while the four-year-olds performed above chance overall, they had only a fragile understanding of the complex sentences. The five-year-olds, in contrast, showed a much better understanding of sentences in iconic clause-order, and of *before*-sentences overall. These findings thus support hypotheses 1 and 2 from the semantic account, (see section 2.3 above), but not hypotheses 3, 4, 5, and 6 from the syntactic and the frequency-based account, respectively.

In the next section, we now turn to the possible role of memory to test the prediction made by the capacity-constrained account (hypothesis 7). In addition, we investigate if individual variation in general language ability and/or executive function is related to complex sentence comprehension, and, if so, if it can explain any additional variance in the children's performance.

3.4. Other Tasks

We first present descriptive statistics for all other tests that were administered. We then test if any of the scores in the memory, language, and executive function tasks are significantly (and with at least substantial evidence) correlated with mean accuracy and/or mean RTs. Those scores that are significantly, and with substantial evidence, correlated with these overall measures are then entered into the optimal statistical models obtained in the analyses above (see section 3.2).

3.4.1. Descriptive Statistics

3.4.1.1. Standardised language and memory tasks

The means and standard deviations of the standardised scores for the CELF and ERB sub-tasks for both age groups are presented in Table 11.

AgeGroup	4		5	
Task	**Mean**	**SD**	**Mean**	**SD**
CELF Linguistic Concepts	10.3	2.2	10.1	2.3
CELF Sentence Structure	10.9	2.9	9.6	3.1
ERB Preschool Repetition Test	101.8	13.4	104.4	14.2
ERB Sentence Imitation Test	95.8	11.6	97.1	11.2

Table 11: Means and standard deviations (SD) of the standardised scores for the CELF and ERB sub-tasks for four-year-olds and five-year-olds.

The means and standard deviations indicate that each group was performing at an age-appropriate level in all of the tasks.

3.4.1.2. Executive function tasks

On the Day/Night task, out of a maximum of 12 correct trials, the mean in the four-year-old group was 11.3 correct responses (SD = 4.3), and 12 (SD = 4.1) in the five-year-old group. In the post-switch phase of the DCCS task, where a maximum of six correct trials are possible, four-year-olds achieved on average 3.6 correct (SD = 2.7), and five-year-olds 4.4 (SD = 2.4). It should be noted, however, that the means are not necessarily informative, because the distribution tends to be bi-modal – children get all trials either wrong or right – which was also the case here. While the four-year-olds were approximately split between 0 and 6 correct responses, the majority of the five-year-olds got all trials correct (see Figure D 1 in Appendix D).

3.4.1.3. Causality task

In both age groups, the mode for correct trials was four (the maximum number of correct trials) indicating that the children showed an understanding of the temporal priority principle.

3.4.2. Correlations with mean accuracy and mean RT

We tested correlations between the z-scores of the language, memory, and executive function tasks and mean accuracy and mean RT scores using standard correlations and Bayesian correlations. The results (tables and corresponding scatterplots) can be found in Appendix D.

Of the six tasks, five were significantly positively correlated with mean accuracy: the CELF Linguistic Concepts score, the CELF Sentence Structure score, the ERB Preschool Repetition test score, the ERB Sentence Imitation test score, and the DCCS post-switch test score. Only the Day/Night score was not significantly correlated with mean accuracy. The Bayes factors obtained through the Bayesian correlation indicate that there was extreme evidence for a correlation with the CELF Linguistic Concepts score, substantial evidence for a correlation with the CELF Sentence Structure test score, and strong evidence for a correlation with the ERB Sentence Imitation. For the DCCS post-switch score, there was only anecdotal evidence for a positive correlation, while there was anecdotal evidence for no correlation between mean accuracy and the ERB Preschool Repetition score, and strong evidence for no correlation between mean accuracy and the Day/Night task score. Overall then, children who scored higher on one of the memory tasks ERB Sentence Imitation) and the standardised language tests (CELF Linguistic Concepts, CELF Sentence Structure) showed better comprehension in the connective comprehension task than children who scored lower.

Three test scores were significantly negatively correlated with response times: the DCCS post-switch phase score, the CELF Linguistic Concepts test score, and the CELF Sentence Structure test score. However, there was strong evidence only for the correlation with the Linguistic Concepts score. The evidence for the correlation with the CELF Sentence structure test score and the DCCS post-switch phase score were only anecdotal. In addition, there was substantial evidence for the lack of a correlation between mean RTs and the Day/Night test scores, and the ERB Preschool Repetition test score. Thus, overall, only the CELF Linguistic Concepts score was strongly negatively correlated with the speed of responses, that is, higher CELF scores were correlated with faster response times.

3.4.3. INFLUENCE ON ACCURACY AND MEAN RT

On the basis of the results of the correlation tests, the CELF Linguistic Concepts score and the two ERB scores (Preschool Repetition and Sentence Imitation), which serve as indicators for working memory, were entered into the optimal model for the prediction of accuracy in the connective comprehension task (see section 3.2.1). Recall that the capacity-constrained account predicts that memory capacity should make an independent contribution to children's performance in the comprehension experiment. Similarly, the CELF Linguistic Concepts score was added to the optimal model for the prediction of response times in the connective comprehension task (see section 3.2.2).

Of the three predictors added to the Accuracy model, only one remained significant and was kept in the model: the CELF Linguistic Concepts score (see Table 12). However, the more complex models that included these additional factors did not converge, a problem that has been noted for mixed-effect models that have multi-level factors (Eager & Roy, 2017). The Bayesian analysis, which did not suffer from non-convergence problems, suggested that the data was 1.5 times more likely under the original model than under the model that included the CELF Linguistic Concepts score ("anecdotal evidence"), and about 23 times more likely under the original model than under the one that included the two memory-related scores, ERB PSRep and ERB Sentence Imitation ("strong evidence"). (For a visualisation, see Figure D 4 in Appendix D.)

	Estimate	Std Error	z value	Pr(>lzl)
(Intercept)	0.14	0.16	0.86	.39
AgeGroup5	-0.32	0.22	-1.42	.16
Typebefore	-0.07	0.19	-0.38	.70
Typebecause	0.22	0.21	1.03	.30
Typeif	0.13	0.21	0.62	.53
ClauseOrdersub-main	0.03	0.19	0.15	.88
scale(LingCon)	**0.15**	**0.06**	**2.45**	**< .01**
AgeGroup5:Typebefore	**1.38**	**0.29**	**4.70**	**< .0001**
AgeGroup5:Typebecause	0.16	0.28	0.56	.58
AgeGroup5:Typeif	-0.09	0.28	-0.34	.74
AgeGroup5:ClauseOrdersub-main	**0.85**	**0.29**	**2.96**	**< .01**
AgeGroup4:Typebefore: ClauseOrdersub-main	0.21	0.27	0.75	.45
AgeGroup5:Typebefore: ClauseOrdersub-main	**-1.35**	**0.31**	**-4.39**	**< .0001**
AgeGroup4:Typebecause: ClauseOrdersub-main	-0.01	0.28	-0.04	.97
AgeGroup5:Typebecause: ClauseOrdersub-main	-0.10	0.30	-0.32	.75
AgeGroup4:Typeif:Clause Ordersub-main	-0.12	0.28	-0.43	.67
AgeGroup5:Typeif: ClauseOrdersub-main	0.09	0.30	0.29	.77

Table 12: Summary of Generalized Linear Mixed Effects Model for the log odds for accuracy responses: effects and Interactions of AgeGroup, Type, and ClauseOrder. The reference levels are for AgeGroup: 4years, for Type: after, and for ClauseOrder: main-subordinate. Significant effects are highlighted in bold font. Note that because Type has four levels, this table shows the results only for the model with "after" as the reference level. The summaries of the models with the other three connectives as reference level showing the same effects can be found in Appendix D.

Standardised memory or language ability scores thus did not explain any additional variation in the accuracy data, over and above the variation that was explained by the interaction of the experimental factors AgeGroup, Type, and ClauseOrder.

For response times, the CELF Linguistic Concepts score was a significant predictor. Children who scored higher on the language test had significantly shorter response times than children who scored lower (see Table 13), suggesting that there may be an independent contribution of general language ability to response times, although the data was about 1.8 times more likely under the Bayesian model without this additional predictor than under the one that included it, which suggests the contribution of the CELF scores to variation in reaction times may be relatively small.

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	4297.39	356.91	109.85	11.90	< 2e^{-16}
AgeGroup5	**-1252.22**	**470.63**	**100.28**	**-2.60**	**< .05**
Typebefore	81.67	218.70	72.78	0.37	.71
Typebecause	**1037.84**	**325.82**	**45.20**	**3.18**	**< .01**
Typeif	**1217.46**	**364.26**	**61.46**	**3.34**	**< .01**
scale(LingCon)	**-487.17**	**225.62**	**100.90**	**-2.11**	**< .05**

Table 13: Summary of Linear Mixed Effects Model for response times (children): effects of Type and Linguistic Concepts scores. The reference level is "after". Significant effects are highlighted in bold font. Note that because Type has four levels, this table shows the results only for "after" as the reference level. The summaries of the models with the other three connectives as reference level showing the same effects can be found in Appendix C.

In summary, although several test scores were correlated with task performance (positively with mean accuracy, negatively with mean response times), none of those predicted any additional variance after accounting for the influence of the experimental factors. In particular, we did not find any evidence for an independent contribution of memory to performance in the connective comprehension task, disconfirming hypothesis 7.

4. Discussion

The aim of this study was to test hypotheses predicted by four different accounts regarding children's processing of complex sentences with the connectives *after*, *before*, *because*, and *if*. In what follows, we first argue that the data support the semantic account best. In the light of the results, we then go on to consider in more detail the role of semantic complexity on the one hand and input frequency on the other. Next we address the production-comprehension asymmetry suggested by our data, before discussing what the results say about the role of individual differences generally, and short-term memory in particular, in language comprehension. In the final part of the discussion, we lay out what it takes to construct a coherent mental model from complex sentences, relating the present research to the wider context of temporal-causal reasoning and the relationship between language and cognitive development.

Iconicity as the key factor in complex sentence comprehension

The children's performance in terms of accuracy is mostly consistent with Clark's (1971) semantic account. The five-year-old children showed a better understanding of sentences in which the order of events in the sentence matched the order of events in the real world (iconic sentences). In addition, they showed better comprehension of *before*-sentences compared to *after*-sentences, and in fact also compared to *because*- and *if*-sentences. Four-year-olds, in contrast, while being above chance overall, showed only a very limited understanding of complex sentences. Our results add to the growing body of evidence that children expect that language directly maps onto the events in the real world, and experience comprehension problems when this is not the case (Blything & Cain, 2016; Blything et al., 2015; Emerson, 1979; Feagans, 1980; French & Brown, 1977; Stevenson & Pollitt, 1987;

Trosborg, 1982). Importantly, our study is the first one to extend this finding to both *because*- and *if*-sentences, suggesting that this is a general principle in children's processing of complex sentences, rather than one that is only employed with temporal clauses. It should be noted, however, that while the error-rates for non-iconic sentences were higher than those for iconic sentences, children did not consistently misinterpret non-iconic sentences as iconic; with the exception of *before*-sentences (which we discuss next), performance was at chance. This may indicate that children find non-iconic sentences un-interpretable, which leads them to choose randomly between two options, rather than imposing an iconic interpretation on every sentence.

Semantic complexity vs. input frequency

Also in support of Clark's semantic account, we found a clear facilitative effect for *before*-sentences, in both clause orders. However, we suggest that this is not due to differences in semantic features, but rather due to a confluence of factors, including frequency and syntactic form. Were it the case that children initially interpreted *after*-sentences as *before*-sentences, as suggested by Clark, they should have performed much worse on *after*-sentences than they did. Instead, these results could suggest that *before* has advantages over *after* in terms of both its semantic transparency, and how often it is used as a connective. Although both *before* and *after* are used more often in other constructions than as temporal connectives, the meaning of *before* is always either spatial ("to appear before the court") or temporal, with clear similarities between the two. The meaning of *after*, however, is often more opaque, as for example in phrasal verbs ("to look after", "to inquire after"). In addition, *before* is used in other constructions only about 1.5 times more often than as a temporal connective in complex sentences, whereas *after* occurs more than four times more often in other constructions in both adult written and spoken language, (Leech, Rayson, & Wilson, 2014), and in child-directed speech (De Ruiter et al., 2017). In other words, *before* has a more consistent form-meaning mapping. For the parser, this means that there is more uncertainty attached to *after* with respect to the construction that is currently being processed, and as a consequence a higher chance of misanalysis. Children's superior performance with iconic *before*-sentences can then be explained by the fact that these combine a lower-uncertainty word (*before*) with an iconic clause order that is main-subordinate, unlike the other three connectives. Our results show clearly that syntactic form in terms of the distance between the subordinator and its resolution is not the determining factor in children's processing of complex sentences, contrary to the syntactic account's prediction. However, in combination with a more consistent form-meaning mapping and iconicity, the shorter recognition domain of the main-subordinate clause order may give iconic *before*-sentences an "edge" over the other sentence types. Iconic *before*-sentences are the only sentences that can be processed incrementally, without re-analysis. We are currently testing the hypothesis that a more consistent form-meaning mapping makes *before*-sentences easier for English children by conducting the same experiment in a language that is similar syntactically, but has different relative frequencies for using the different words as connectives: German. If the hypothesis is correct, the advantage of (non-iconic) *before*-type-sentences should then disappear. If, on the other hand, the effect persists, this would support a semantic explanation along the lines of Clark (1971).

If (relative) frequency does have some role to play in complex sentence comprehension after all, then the question is: Why were children in our experiment not better at comprehending *because*- and *if*-clauses, which are much more frequent in English than *after*- and *before*-sentences? In the present study, the children showed in the causality task that they did understand that causes must precede effects, and the older age-group showed an understanding of *because*- and *if*-sentences in iconic order. But despite understanding some aspects of causality, performance was relatively low. Furthermore, children of both age groups were significantly slower in responding to *because*- and *if*-sentences compared to *after*- and *before*-sentences.

One possible explanation is due to the sentences' higher semantic complexity. Understanding isolated *because*- and *if*-sentences requires an understanding of both temporality and causality, purely through language, whereas *before*- and *after*-sentences rely on temporality only. Furthermore, causality may be semantically more complex than temporality: it has been observed that in production, children use the connective *and* to express semantic relations in the order of additive < temporal < causal < adversative (L Bloom, Lahey, Hood, Lifter, & Fiess, 1980), which have been said to be of increasing semantic complexity, following the notion of cumulative complexity introduced by Brown (1973). But if the cumulative complexity assumption holds also for comprehension, it remains unclear why there was no difference in accuracy between the semantically simpler *after*-sentences on the one hand, and the semantically more complex *because*- and *if*-sentences on the other. Interestingly, the response time data are in line with the assumption of cumulative complexity: Responses to *because*- and *if*-sentences were slower than to *after*- and *before*-sentences. This suggests that processing two clauses that are causally linked takes longer than processing clauses that are only temporally linked. There is thus an interesting disconnect between the accuracy data, which showed an advantage for iconic sentences, and for *before*-sentences in general, and the RT data, which showed an advantage for temporal clauses. It is possible that children perceive temporal sentences to be easier (and thus react more quickly), even if their actual levels of accuracy indicate comprehension difficulties, at least for (non-iconic) *after*-sentences. Processing causal sentences may take more time, but it does not necessarily lead to more errors.

Production-comprehension asymmetry

An argument against the cumulative complexity account as an explanation is that children also start producing *because*- and *if*-sentences before they start producing *after*- and *before*-sentences (e.g., Diessel, 2004), suggesting that they find *because*- and *if*-sentences easier. Production-comprehension asymmetries raise interesting questions in language acquisition research, and different accounts have been put forward. (see e.g., Grimm, Müller, Hamann, & Ruigendijk, 2011). Here we suggest two possible explanations for this mismatch. First, it may be that producing *because*- and *if*-sentences in natural interaction puts different demands on children than comprehending them in an experiment. In spontaneous production, children go from intended meaning to form, all within a supporting linguistic and non-linguistic context (usually the here-and-now). They already know what the relation is between two events they want to express. They can also avoid more complex forms and use alternative strategies (e.g., stringing clauses together using "and then" to express temporal order, instead of an *after-/before*-sentence). In comprehension, and in particular in experiments that do not provide any additional context, children need to rely purely on form to understand the meaning (we discuss the requirements for constructing meaning below). Second, it may be that children are less familiar with *because*- and *if*-sentences being used to express physical causality. Recall that in everyday conversation, speakers use *because*-clauses primarily to give reasons for a preceding speech act ("You can't have sweets now because we're having dinner soon"), and *if*-clauses often provide a conceptual framework for a larger chunk of discourse ("If I ever win the lottery, I have plenty ideas of what to do with the money."). On the other hand, both experimental and observational studies have found that at least Dutch children are able to express content-type causality from three years onwards, suggesting this domain is not uncommon for young children (Evers-Vermeul & Sanders, 2011). Future studies should investigate how providing more context or using other types of causality affects children's comprehension of causal sentences.

Individual differences and memory

Turning now to the role of individual differences, we found that the accuracy data and the RT data showed similar patterns with respect to their relationship with individual measures of language ability, memory, and executive function (inhibition). Children with higher scores on these tasks achieved higher accuracy in the comprehension task, and responded more quickly. However, these factors did not explain any variation in performance after effects of age, type of sentence, and clause order were accounted for. In particular, we did not find any evidence for an independent contribution of memory, contrary to the predictions made by the capacity-constrained account. Note that not only did we not find any significant effect of memory; using a Bayesian approach, we found strong evidence *against* the role of memory and other measures in the models. It is possible that our measures (word- and non-word repetition and sentence imitation) did not capture the type of memory that is central to complex sentence comprehension. Blything et al. (2015) and Blything and Cain (2016), who observed a memory effect, used a digit-span task. However, in view of the fact that the researchers who originally proposed the memory capacity-constrained account measured memory capacity using reading span (Just & Carpenter, 1992), we believe that with children, sentence imitation (with sentences of increasing length) is a comparable measure. Against this background, our results do not provide evidence for a significant role of individual differences in memory, executive function, and general language ability in complex sentence comprehension. This contrasts with other studies that have found that variability in aspects such as working memory or executive function is associated with different language outcomes, even after controlling for age (e.g., Blything & Cain, 2016; White, Alexander, & Greenfield, 2017), but our findings are far from uncommon, as the picture is rather mixed (see Kidd, 2013 for a critical review of the role of working memory). Overall, our findings suggest that the ability to construct a coherent mental model from isolated complex sentences is not just a competence emerging from a combination of general language ability, memory, and executive function, but a distinct construct that cannot be captured with standardised tests.

What is this construct and how does it develop over time? We first discuss our results in relation to previous studies, before connecting them to the wider context of the development of temporal-causal reasoning, and the relationship between language and cognitive development.

Temporal-causal reasoning and the construction of mental event representations

In our data, four-year-olds showed only a rudimentary ability to process complex sentences in isolation, whereas the five-year-olds showed a more robust – albeit still incomplete – understanding. For *before* and *after*, this contrasts with some previous studies, which found above-chance performance at a slightly younger age, between three and four years (e.g., Blything et al., 2015). We attribute this difference to the fact that the task required that the children consider two explicit alternatives ("story" A and "story" B) before making a selection. As we discuss below, this requires that the listener have a stable mental representation of the events, which she can handle flexibly to reason about temporal and causal relations between them. For *because* and *if*, our findings are more in line with those of Amidon (1976), who found above-chance performance in her youngest age group (five years), and not with those of Emerson (1979) and Emerson and Gekoski (1980), who found children to comprehend *because*- and *if*-sentences only around the age of eight years.

Research on children's capacity to reason (non-linguistically) about temporal and causal relations events using search and planning tasks has found that flexible temporal–causal reasoning develops around the age of five or six years (Lohse, Kalitschke, Ruthmann, & Rakoczy, 2015; e.g., McCormack & Hanley, 2011). The basic logic of the tasks is that participants need to mentally reconstruct or pre-construct a sequence of causally linked events in order to correctly infer a present or anticipated future state of the world (e.g., an object's location). While four-year-olds usually do not have problems understanding the temporal priority principle (Rankin & McCormack, 2013) – as in the present study –, it appears that they cannot perform in these search and planning tasks unless under specific conditions, indicating that they lack the capacity to reason flexibly about temporal-causal relations. Specifically, younger children seem to be able to perform this task only when it refers to past events, but not when they have to mentally construct a sequence of events themselves to make inferences (McCormack & Hanley, 2011). Furthermore, younger children appear to require visible, positive evidence (e.g., a clear sign that an object had been used in a particular location) to infer a state of events (e.g., that the object must have been lost after it was used in that location). Older children, in contrast, can also use the absence of evidence to perform inferences (i.e., use counterfactual reasoning; Lohse et al., 2015).

How could this background help explain the difference between the current findings and those of Blything et al. (2015), who found that their youngest age group (three-to-four-year-olds) performed better with *before*- and *after*-sentences than the four-year-olds in the present study? In Blything et al.'s study, children watched short animated clips of the actions of both clauses of the complex sentence (e.g., eating a hotdog, putting shoes on) successively next to each other, which ended in a freeze frame. They then heard the prompt "Listen carefully and touch the thing Tom/Sue did first", followed by the sentence (e.g., "Before he ate the burger, he put on the sandals"). In contrast, in the present study, children first heard the prompt, followed by the sentence (e.g., "After she paints the old fence, she hoovers the house"), and then saw the two picture stories. The children in Blything et al.'s study were aware that they had to pay attention only to what happened first, and they knew what the two possible actions were before even hearing the sentence. The children in the present study had to first construct a mental representation of the chain of events from language only, without any initial visual support, and then needed to check this model against two possible laid out sequences. The research on temporal-causal reasoning outlined above suggests that creating a mental sequence "from scratch" may be challenging for four-year-olds, so we would expect those representations to be more fragile than those that are supported visually from the start, and may not yet be stable enough to reason about them in order to make a selection on the screen (e.g., “if this is what happens, then the story at the top must be the right one”). We suggest that the task used in the present study is actually a closer match to what listeners typically have to do: construct a mental model from the speech input alone, and use that model subsequently, for example to make a decision (e.g., “Before you do your homework, put your clothes in the laundry basket” – what needs to happen now?).

The relationship between language and cognitive development

An important question arising from these different strands of research concerns the mutual influence of language and cognition. Is it the development of temporal-causal reasoning capacities that allows children to understand complex sentences describing chains of events in different ways (iconic and non-iconic)? Or is it children’s situated language experience that leads them to develop more flexible representations of events? For example, a child may encounter a non-iconic sentence in a situation where the real-world context makes it clear what the order of events is (“Before you go to bed you need to brush your teeth”), which enables her to understand that language can describe events in non-iconic ways, which in turn leads to a more abstract and flexible understanding of how two (or more) events are linked. It seems likely that as in other areas of language and cognitive development (e.g. complex complement clauses and theory of mind, De Villiers, 2007), a bidirectional relationship exists with developments in each domain supporting the other.

In the context of causal reasoning, it is interesting to note that the few errors that the adults made in the present study occurred almost exclusively (eight out of eleven) with one item. The test sentence was “If/because she dives in the pool, she feels really warm”, and the correct story was one showing the protagonist diving into a heated pool in a wintery landscape outside and enjoying the warmth, whereas the foil sequence shows her standing in the sun in the summer and then diving into a (cold) pool. It appears that several adult participants interpreted the sentence in an epistemic way, in the sense of "If she dives in the pool then that must mean that she's feeling warm", which makes the foil the better match. This item did not stand out from the other items in the children's data, which suggests that this epistemic interpretation may not yet have been open to them. This would be in line with corpus studies of English, French, and Dutch child language, which have found that subjective causal relations appear later than objective relations (e.g., Evers-Vermeul & Sanders, 2011; Zufferey, Mak, & Sanders, 2015).

The five-year-olds in the present study were still far from adult-like in their performance. It is clear that complex sentence comprehension must undergo substantial development throughout the school years. School education, and literacy training in particular, is likely to contribute to this development. Children are exposed to written texts and taught to pay attention to elements that link clauses and sentences with each other in order to understand the meaning of a text. This will also impact their spoken language comprehension. Furthermore, children will develop their understanding (and production) of other forms of causal language, in particular epistemic language. At this point it is still unclear what the role of the input (either spoken or written) may be in children's development of different forms of causal language.

This study investigated the role of syntax, semantics, frequency, and working memory in the comprehension of complex sentences involving adverbial clauses. To limit the availability of additional cues to meaning and therefore provide a relatively pure test, sentences were deliberately presented with minimal contextual support. Of course, in reality, complex sentences are typically used in discourse, and thus another question concerns how their processing is affected by information structure, or discourse pragmatics. It has been found that adult listeners find sentences in which given information precedes new information easier to process (Haviland & Clark, 1974) and there is an indication that young children (three to five years) prefer a given-before-new order in *when*-sentences containing a main and subordinate clause (Junge, Theakston, & Lieven, 2015). An interesting avenue for future studies would be to explore how information structure affects children's comprehension of different types of complex sentences, and to what extent such an effect may interact with the effect of iconicity that we found in our study.

5. Summary

In this paper, we provide the most comprehensive experimental study to date to evaluate four theoretical models of the factors underpinning children's abilities to comprehend complex sentences containing adverbial clauses. We found that children's comprehension was strongly influenced by semantic factors – the iconicity of the event-to-language mappings – and their response times were influenced by the type of relation expressed (temporal vs. causal). We found that neither input frequency (frequency-based account), nor clause order (syntax account) or working memory (capacity-constrained account) provided a good fit to the data. Our findings thus contribute to the development of more sophisticated models of sentence processing to apply through acquisition and into adulthood. Although the stimuli used in the present study were deliberately designed to be challenging, we would argue that they reflect the demands placed on children in everyday life, especially in academic contexts. We conclude that models of linguistic processing and representation must take into account how children's emerging linguistic understanding interacts with developments in other cognitive domains such as their ability to construct mental models and reason flexibly about them.

Bibliography

Aben, B., Stapert, S., & Blokland, A. (2012). About the distinction between working memory and short-term memory. *Frontiers in Psychology*, *3*.

Amidon, A. (1976). Children's understanding of sentences with contingent relations: Why are temporal and conditional connectives so difficult? *Journal of Experimental Child Psychology*, *22*(3), 423–437. https://doi.org/10.1016/0022-0965(76)90106-5

Amidon, A., & Carey, P. (1972). Why five-year-olds cannot understand before and after. *Journal of Verbal Learning and Verbal Behavior*, *11*(4), 417–423. https://doi.org/10.1016/S0022-5371(72)80022-7

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. Retrieved from http://arxiv.org/abs/1506.04967

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bloom, L., & Capatides, J. B. (1987). Sources of meaning in the acquisition of complex syntax: The sample case of causality. *Journal of Experimental Child Psychology*, *43*(1), 112–128. https://doi.org/10.1016/0022-0965(87)90054-3

Bloom, L., Lahey, M., Hood, L., Lifter, K., & Fiess, K. (1980). Complex sentences: acquisition of syntactic connectives and the semantic relations they encode. *Journal of Child Language*, *7*(2), 235–261. https://doi.org/10.1017/S0305000900002610

Blything, L. P., & Cain, K. (2016). Children’s processing and comprehension of complex sentences containing temporal connectives: The influence of memory on the time course of accurate responses. *Developmental Psychology*, *52*(10), 1517–1529. https://doi.org/10.1037/dev0000201

Blything, L. P., Davies, R., & Cain, K. (2015). Young children’s comprehension of temporal relations in complex sentences: The influence of memory on performance. *Child Development*, *86*(6), 1922–1934.

Boersma, P., & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Retrieved from http://www.praat.org/

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

Brown, R. (1973). *A first language: The early stages.* Harvard University Press.

Carni, E., & French, L. (1984). The acquisition of before and after reconsidered: what develops? *Journal of Experimental Child Psychology*, *37*, 394–403. https://doi.org/10.1016/0022-0965(84)90011-0

Clark, E. V. (1971). On the acquisition of the meaning of “before” and “after.” *Journal of Verbal Learning and Verbal Behavior*, *10*, 266–275.

Clark, H. H., & Clark, E. V. (1968). Semantic distinctions and memory for complex sentences. *The Quarterly Journal of Experimental Psychology*, *20*(2), 129–138. https://doi.org/10.1080/14640746808400141Corrigan, R. (1975). A scalogram analysis of the development of the use and comprehension of “because” in children. *Child Development*, *46*, 195–201.

Dancygier, B., & Sweetser, E. (2000). Constructions with if, since, and because: Causality, epistemic stance, and clause order. In *Mental spaces in grammar: Conditional constructions* (pp. 111–142).

De Ruiter, L. E., & Theakston, A. L. (2017). First language acquisition. In B. Dancygier (Ed.), *Cambridge Handbook of Cognitive Linguistics* (pp. 59–72). Cambridge University Press.

De Ruiter, L. E., Theakston, A. L., Brandt, S., & Lieven, E. V. M. (2017). *The relationship between parental input and children's spontaneous use of adverbial clauses containing after, before, because, and if*. Poster presented at the 14th International Congress for the Study of Child Language (IASCL), July 17-21, Lyon, France.

De Villiers, J. (2007). The interface of language and theory of mind. *Lingua*, *117*(11), 1858–1878.

Dienes, Z. (2011). Bayesian versus orthodox statistics: which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(July), 1–17. https://doi.org/10.3389/fpsyg.2014.00781

Diessel, H. (2001). The ordering distribution of main and adverbial clauses: a typological study. *Language*, *77*(3), 433–455. https://doi.org/10.1353/lan.2001.0152

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge: Cambridge University Press.

Diessel, H. (2005). Competing motivations for the ordering of main and adverbial clauses. *Linguistics*, *43*(3), 449–470. https://doi.org/10.1515/ling.2005.43.3.449

Diessel, H. (2008). Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics*, *19*(3), 465–490. https://doi.org/10.1515/COGL.2008.018

Diessel, H., & Hetterle, K. (2011). Causal clauses: a cross-linguistic investigation of their structure, meaning, and use. In P. Siemund (Ed.), *Linguistic universals and language variation* (pp. 21–52). Mouton de Gruyter. https://doi.org/10.4324/9780203209493_Introduction

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, *81*(4), 882–906. https://doi.org/10.1353/lan.2005.0169

Eager, C., & Roy, J. (2017). Mixed effects models are sometimes terrible. Retrieved from http://arxiv.org/abs/1701.04858

Emerson, H. F. (1979). Children's comprehension of "because" in reversible and non-reversible sentences. *Journal of Child Language*, *6*, 279–300. https://doi.org/10.1017/S0305000900002300

Emerson, H. F. (1980). Children's judgements of correct and reversed sentences with "if". *Journal of Child Language*, *7*(7), 137–155. https://doi.org/10.1017/S0305000900007078

Emerson, H. F., & Gekoski, W. L. (1980). Development of comprehension of sentences with "because" or "if." *Journal of Experimental Child Psychology*, *29*(2), 202–224. https://doi.org/10.1016/0022-0965(80)90016-8

Evers-Vermeul, J., & Sanders, T. (2011). Discovering domains - On the acquisition of causal connectives. *Journal of Pragmatics*, *43*(6), 1645–1662. https://doi.org/10.1016/j.pragma.2010.11.015

Feagans, L. (1980). Children's understanding of some temporal terms denoting order, duration, and simultaneity. *Journal of Psycholinguistic Research*, *9*(1), 41–57.

https://doi.org/10.1007/BF01067301

Ford, C. E., & Thompson, S. A. (1986). Conditionals in discourse: A text-based study from English. *On Conditionals*, 353–372.

French, L. A., & Brown, A. L. (1977). Comprehension of “before” and “after” in logical and arbitrary sequences. *Journal of Child Language*, *4*(2), 247–256. https://doi.org/10.1017/S0305000900001641

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: performance of children 3 1/2- 7 years old on a stroop- like day-night test. *Cognition*, *53*(2), 129–153. https://doi.org/10.1016/0010-0277(94)90068-X

Gorrell, P., Crain, S., & Fodor, J. D. (1989). Contextual information and temporal terms. *Journal of Child Language*, *16*(3), 623–632. https://doi.org/10.1017/S0305000900010758

Grimm, A., Müller, A., Hamann, C., & Ruigendijk, E. (2011). *Production-comprehension asymmetries in child language* (Vol. 43). Walter de Gruyter.

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. *Journal of Statistical Software*, *59*(9), 1–30. Retrieved from http://www.jstatsoft.org/v59/i09/

Haviland, S. E., & Clark, H. H. (1974). What’s new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 512–521.

Hawkins, J. A. (1990). A parsing theory of word order universals. *Linguistic Inquiry*, *21*(2), 223–261.

Hawkins, J. A. (1992). Syntactic weight versus information structure in word order variation. In *Informationsstruktur und Grammatik* (pp. 196–219). Springer.

Hawkins, J. A. (1994). *A performance theory of order and constituency* (Cambridge, Vol. 73). Cambridge University Press.

Irwin, J. W. (1980). The effects of explicitness and clause order on the comprehension of reversible causal relationships. *Reading Research Quarterly*, *15*(4), 477–488. Retrieved from http://www.jstor.org/stable/747275

Jeffreys, H. (1961). *Theory of probability* (3rd Edition). Oxford: Clarendon Press.

Johnson, H. L. (1975). The meaning of before and after for preschool children. *Journal of Experimental Child Psychology*, *19*(1), 88–99. https://doi.org/10.1016/0022-0965(75)90151-4

Johnson, H. L., & Chapman, R. S. (1980). Children's judgment and recall of causal connectives: A developmental study of "because," "so," and "and." *Journal of Psycholinguistic Research*, *9*(3), 243–260. https://doi.org/10.1007/BF01067240

Junge, B., Theakston, A. L., & Lieven, E. V. M. (2015). Given–new/new–given? Children's sensitivity to the ordering of information in complex sentences. *Applied Psycholinguistics*, *36*(3), 589–612. https://doi.org/10.1017/S0142716413000350

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Keller-Cohen, D. (1987). Context and strategy in acquiring temporal connectives. *Journal of Psycholinguistic Research*, *16*(2), 165–183. https://doi.org/10.1007/BF01072000

Kidd, E. (2013). The role of verbal working memory in children's sentence comprehension: a critical review. *Topics in Language Disorders*, *33*(3), 208–223. https://doi.org/10.1097/TLD.0b013e31829d623e

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), 37–41. https://doi.org/10.1080/01690960601155284

Kuhn, D., & Phelps, H. (1976). The development of children's comprehension of causal direction. *Child Development*, *47*(1), 248–251. https://doi.org/10.2307/1128307

Kuznetsova, A., Brockhoff, P. B., & Bojesen Christensen, R. H. (2016). lmerTest: Tests in Linear Mixed Effects Models. Retrieved from https://cran.r-project.org/package=lmerTest

Leech, G., Rayson, P., & Wilson, A. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Lohse, K., Kalitschke, T., Ruthmann, K., & Rakoczy, H. (2015). The development of reasoning about the temporal and causal relations among past, present, and future events. *Journal of Experimental Child Psychology*, *138*, 54–70. https://doi.org/10.1016/j.jecp.2015.04.008

Magimairaj, B. M., & Montgomery, J. W. (2012). Children's verbal working memory: role of processing complexity in predicting spoken sentence comprehension. *Journal of Speech, Language, and Hearing Research*, *55*(June), 669–683. https://doi.org/10.1044/1092-4388(2011/11-0111)a

McCormack, T., & Hanley, M. (2011). Children's reasoning about the temporal order of past and future events. *Cognitive Development*, *26*(4), 299–314. https://doi.org/10.1016/j.cogdev.2011.10.001

Montgomery, J. W., Magimairaj, B. M., & O'Malley, M. H. (2008). Role of working memory in typically developing children's complex sentence comprehension. *Journal of*

Psycholinguistic Research, *37*(5), 331–354. https://doi.org/10.1007/s10936-008-9077-z

Morey, R. D., & Rouder, J. N. (2011). Bayes Factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. https://doi.org/10.1037/a0024377

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes Factors for Common Designs. Retrieved from https://cran.r-project.org/package=BayesFactor

Morey, R. D., & Wagenmakers, E. J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121–124. https://doi.org/10.1016/j.spl.2014.05.010

Münte, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, *395*(6697), 71–73. https://doi.org/10.1038/25731

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (2014). BayesMed: Default Bayesian Hypothesis Tests for Correlation, Partial Correlation, and Mediation. Retrieved from http://cran.r-project.org/package=BayesMed

Phillips, N. (2016). yarrr: A Companion to the e-Book “YaRrr!: The Pirate’s Guide to R.” Retrieved from https://cran.r-project.org/package=yarrr

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. In *Mixed-effects models in S and S-PLUS* (pp. 1–29). New York: Springer-Verlag. https://doi.org/10.1007/0-387-22747-4_1

Pyykönen, P., Niemi, J., & Järvikivi, J. (2003). Sentence structure, temporal order and linearity: slow emergence of adult-like syntactic performance in Finnish. *SKY Journal of Linguistics*, (16), 113–138.

R Core Team. (2016). R: A language and environment for statistical computing. Austria.

Rankin, M. L., & McCormack, T. (2013). The temporal priority principle: At what age does this develop. *Frontiers in Psychology*, *4*(MAY). https://doi.org/10.3389/fpsyg.2013.00178

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Seeff-Gabriel, B. K., Chiat, S., & Roy, P. (2008). Early Repetition Battery.

Smith, K. H., & McMahon, L. E. (1970). Understanding order information in sentences: Some recent work at the Bell Laboratories. In W. J. M. Levelt & G. B. Flores d'Arcais (Eds.), *Advances in Psycholinguistics* (pp. 253–279). Amsterdam: North Holland.

Stevenson, R. J., & Pollitt, C. (1987). The acquisition of temporal terms. *Journal of Child Language*, *14*(3), 533–545. https://doi.org/10.1017/S0305000900010278

Sweetser, E. (1990). *From etymology to pragmatics: metaphorical and cultural aspects of semantic structure*. Cambridge University Press. https://doi.org/10.1016/S0378-2166(96)90003-X

Trosborg, A. (1982). Children's comprehension of "before" and "after" reinvestigated. *Journal of Child Language*, *9*(2), 381–402. https://doi.org/10.1017/S0305000900004773

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

White, L. J., Alexander, A., & Greenfield, D. B. (2017). The relationship between executive functioning and language: examining vocabulary, syntax, and language learning in preschoolers attending Head Start. *Journal of Experimental Child Psychology*, *164*, 16–

31. https://doi.org/10.1016/j.jecp.2017.06.010

Wiig, E. H., Secord, W., & Semel, E. M. (2004). CELF preschool 2: Clinical Evaluation of Language Fundamentals preschool. Pearson/PsychCorp.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): a method of assessing executive function in children. *Nature Protocols*, *1*(1), 297–301. https://doi.org/10.1038/nprot.2006.46

Zufferey, S., Mak, W. M., & Sanders, T. J. M. (2015). A cross-linguistic perspective on the acquisition of causal connectives and relations. *International Review of Pragmatics*, *7*, 22–39. https://doi.org/10.1163/18773109-00701002

APPENDIX A

Session	Sentence no.	Sentence List 1	Sentence List 3
1	1	**After** she paints the old fence, she hoovers the house.	**Before** she paints the old fence, she hoovers the house.
	2	**After** he sweeps the new floor, he watches TV.	**Before** he sweeps the new floor, he watches TV.
	3	He drinks some water, **after** he eats a green pear.	He drinks some water, **before** he eats a green pear.
	4	He laughs really hard, **after** he coughs a few times.	He laughs really hard, **before** he coughs a few times.
	5	She hides over there, **after** she runs over here.	She hides over there, **before** she runs over here.
	6	**After** she dances around, she bounces away.	**Before** she dances around, she bounces away.
	7	**Before** he reads his new book, he plays his big drum.	**After** he reads his new book, he plays his big drum.
	8	She takes a hot bath, **before** she draws a picture.	She takes a hot bath, **after** she draws a picture.
	9	She breaks her small train, **before** she builds a tower.	She breaks her small train, **after** she builds a tower.
	10	She hops up and down, **before** she crawls on the floor.	She hops up and down, **after** she crawls on the floor.
	11	**Before** he shouts out loudly, he drives away fast.	**After** he shouts out loudly, he drives away fast.
	12	**Before** he waves happily, he swims on his back.	**After** he waves happily, he swims on his back.
	13	**Because** she bangs her head hard, she closes her eyes.	**If** she bangs her head hard, she closes her eyes.
	14	**Because** he opens the door, he sees the snowman.	**If** he opens the door, he sees the snowman.
	15	He misses the bus, **because** he rides his old bike.	He misses the bus, **if** he rides his old bike.
	16	He cries really hard, **because** he trips suddenly.	He cries really hard, **if** he trips suddenly.
	17	She feels really warm, **because** she dives in the pool.	She feels really warm, **if** she dives in the pool.
	18	**Because** she looks at the sky, she slips to the ground.	**If** she looks at the sky, she slips to the ground.
	19	**If** he sings a happy song, he wins a nice cup.	**Because** he sings a happy song, he wins a nice cup.
	20	She finds her other shoe, **if** she cuts the long grass.	She finds her other shoe, **because** she cuts the long grass.
	21	She hears the doorbell, **if** she presses the button.	She hears the doorbell, **because** she presses the button.
	22	She wakes up in the night, **if** she talks to herself.	She wakes up in the night, **because** she talks to herself.
	23	**If** he sits down in his chair, he gets very bored.	**Because** he sits down in his chair, he gets very bored.
	24	**If** he sneezes lots of times, he falls in the field.	**Because** he sneezes lots of times, he falls in the field.

2	1	She hoovers the house, **after** she paints the old fence.	She hoovers the house, **before** she paints the old fence.
	2	He watches TV, **after** he sweeps the new floor.	He watches TV, **before** he sweeps the new floor.
	3	**After** he eats a green pear, he drinks some water.	**Before** he eats a green pear, he drinks some water.
	4	**After** he coughs a few times, he laughs really hard.	**Before** he coughs a few times, he laughs really hard.
	5	**After** she runs over here, she hides over there	**Before** she runs over here, she hides over there
	6	She bounces away, **after** she dances around.	She bounces away, **before** she dances around.
	7	He plays his big drum, **before** he reads his new book.	He plays his big drum, **after** he reads his new book.
	8	**Before** she draws a picture, she takes a hot bath.	**After** she draws a picture, she takes a hot bath.
	9	**Before** she builds a tower, she breaks her small train.	**After** she builds a tower, she breaks her small train.
	10	**Before** she crawls on the floor, she hops up and down.	**After** she crawls on the floor, she hops up and down.
	11	He drives away fast, **before** he shouts out loudly.	He drives away fast, **after** he shouts out loudly.
	12	He swims on his back, **before** he waves happily.	He swims on his back, **after** he waves happily.
	13	She closes her eyes, **because** she bangs her head hard.	She closes her eyes, **if** she bangs her head hard.
	14	He sees the snowman, **because** he opens the door.	He sees the snowman, **if** he opens the door.
	15	**Because** he rides his old bike, he misses the bus.	**If** he rides his old bike, he misses the bus.
	16	**Because** he trips suddenly, he cries really hard.	**If** he trips suddenly, he cries really hard.
	17	**Because** she dives in the pool, she feels really warm.	**If** she dives in the pool, she feels really warm.
	18	She slips to the ground, **because** she looks at the sky.	She slips to the ground, **if** she looks at the sky.
	19	He wins a nice cup, **if** he sings a happy song.	He wins a nice cup, **because** he sings a happy song.
	20	**If** she cuts the long grass, she finds her other shoe.	**Because** she cuts the long grass, she finds her other shoe.
	21	**If** she presses the button, she hears the doorbell.	**Because** she presses the button, she hears the doorbell.
	22	**If** she talks to herself, she wakes up in the night.	**Because** she talks to herself, she wakes up in the night.
	23	He gets very bored, **if** he sits down in his chair.	He gets very bored, **because** he sits down in his chair.
	24	He falls in the field, **if** he sneezes lots of times.	He falls in the field, **because** he sneezes lots of times.

Table A 1: Experimental sentences for the experimental Lists 1 and 3. Note that in List 3, all after-sentences from List 1 have been changed to before-sentences, and vice versa. In the same way, all because-sentences from List 1 were changed to if-sentences in List 3, and vice versa. Experimental lists 2 and 4 were created by swapping session 1 and 2 of List 1 and List 3, respectively.

APPENDIX B

Fixed effects	Estimate	Std. Error	z value	Pr(>lzl)
(Intercept)	-0.01	0.16	-0.04	097
AgeGroup5	**1.21**	**0.23**	**5.19**	**< .0001**
Typeafter	0.08	0.19	0.39	.7
Typebecause	0.29	0.21	1.38	.17
Typeif	0.21	0.21	0.97	.33
ClauseOrdersub-main	0.24	0.19	1.21	.22
AgeGroup5:Typeafter	**-1.38**	**0.29**	**-4.71**	**< .0001**
AgeGroup5:Typebecause	**-1.22**	**0.3**	**-4.15**	**< .0001**
AgeGroup5:Typeif	**-1.48**	**0.29**	**-5.01**	**< .0001**
AgeGroup5:ClauseOrdersub-main	**-0.71**	**0.3**	**-2.4**	**< .05**
AgeGroup4:Typeafter:ClauseOrdersub-main	-0.21	0.27	-0.76	0.45
AgeGroup5:Typeafter:ClauseOrdersub-main	**1.35**	**0.31**	**4.39**	**< .0001**
AgeGroup4:Typebecause:ClauseOrdersub-main	-0.22	0.28	-0.79	.43
AgeGroup5:Typebecause:ClauseOrdersub-main	**1.26**	**0.31**	**4.02**	**< .001**
AgeGroup4:Typeif:ClauseOrdersub-main	-0.32	0.28	-1.18	0 24
AgeGroup5:Typeif:ClauseOrdersub-main	**1.44**	**0.31**	**4.64**	**<. 0001**

Table B 1: Summary of Generalized Linear Mixed Effects Model for the Accuracy responses: effects and Interactions of AgeGroup, Type, and ClauseOrder. The reference levels are for AgeGroup: 4years, for Type: before, and for ClauseOrder: main-subordinate. Significant effects are highlighted in bold font. Note that because before is the reference level (for which the 5-year-olds performed better than the 4-year-olds), this model shows a main effect of AgeGroup, with the 5-year-olds being significantly better than the 4-year-olds.

Fixed effects	Estimate	Std. Error	z value	Pr(>lzl)
(Intercept)	0.29	0.16	1.76	.08
AgeGroup5	-0.01	0.22	-0.06	.96
Typebefore	-0.29	0.21	-1.38	.17
Typeafter	-0.22	0.21	-1.02	.31
Typeif	-0.09	0.20	-0.45	.66
ClauseOrdersub-main	0.02	0.20	0.10	.92
AgeGroup5:Typebefore	**1.22**	**0.30**	**4.14**	**< .0001**
AgeGroup5:Typeafter	-0.16	0.28	-0.57	.57
AgeGroup5:Typeif	-0.25	0.28	-0.89	.38
AgeGroup5:ClauseOrdersub-main	**0.76**	**0.29**	**2.60**	**< .05**
AgeGroup4:Typebefore:ClauseOrdersub-main	0.22	0.28	0.79	.43
AgeGroup5:Typebefore:ClauseOrdersub-main	**-1.26**	**0.31**	**-4.02**	**< .0001**
AgeGroup4:Typeafter:ClauseOrdersub-main	0.01	0.28	0.03	.97
AgeGroup5:Typeafter:ClauseOrdersub-main	0.10	0.30	0.32	.75
AgeGroup4:Typeif:ClauseOrdersub-main	-0.11	0.28	-0.39	.70
AgeGroup5:Typeif:ClauseOrdersub-main	0.18	0.30	0.60	.55

Table B 2: Summary of Generalized Linear Mixed Effects Model for the Accuracy responses: effects and Interactions of AgeGroup, Type, and ClauseOrder. The reference levels are for AgeGroup: 4years, for Type: because, and for ClauseOrder: main-subordinate. Significant effects are highlighted in bold font.

Fixed effects	Estimate	Std. Error	z value	Pr(>lzl)
(Intercept)	0.20	0.16	1.22	.22
AgeGroup5	-0.26	0.22	-1.21	.23
Typebecause	0.09	0.20	0.45	.65
Typebefore	-0.20	0.21	-0.97	.33
Typeafter	-0.13	0.21	-0.61	.54
ClauseOrdersub-main	-0.09	0.20	-0.45	.65
AgeGroup5:Typebecause	0.25	0.28	0.89	.38
AgeGroup5:Typebefore	**1.47**	**0.29**	**5.00**	**<.0001**
AgeGroup5:Typeafter	0.09	0.28	0.32	.75
AgeGroup5:ClauseOrdersub-main	**1.05**	**0.29**	**3.63**	**<.001**
AgeGroup4:Typebecause:ClauseOrdersub-main	0.11	0.28	0.38	.70
AgeGroup5:Typebecause:ClauseOrdersub-main	-0.18	0.30	-0.60	.55
AgeGroup4:Typebefore:ClauseOrdersub-main	0.32	0.28	1.17	.24
AgeGroup5:Typebefore:ClauseOrdersub-main	**-1.44**	**0.31**	**-4.64**	**<.0001**
AgeGroup4:Typeafter:ClauseOrdersub-main	0.12	0.28	0.42	.67
AgeGroup5:Typeafter:ClauseOrdersub-main	-0.09	0.30	-0.29	.77

Table B 3: Summary of Generalized Linear Mixed Effects Model for the Accuracy responses: effects and Interactions of AgeGroup, Type, and ClauseOrder. The reference levels are for AgeGroup: 4years, for Type: if, and for ClauseOrder: main-subordinate. Significant effects are highlighted in bold font.

APPENDIX C

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	4667.62	357.04	88.87	13.073	< $2e^{-16}$
AgeGroup5	1750.53	417.22	65.68	-4.196	$8.34E^{-05}$
Typeafter	-89.33	219.39	136.53	-0.407	.68454
Typebecause	**958.04**	**303.26**	**36.28**	**3.159**	**< .05**
Typeif	**1138.22**	**337.53**	**46.96**	**3.372**	**< .01**

Table C 1: Summary of Linear Mixed Effects Model for response times: effects of AgeGroup and Type. The reference levels are for AgeGroup: 4years, for Type: before. Significant effects are highlighted in bold font.

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	5625.68	385.31	108.72	14.494	< $2e^{-16}$
AgeGroup5	-1750.68	417.23	99.86	-4.095	$8.58e^{-05}$
Typebefore	**-1047.37**	**327.32**	**45.44**	**-3.193**	**< .01**
Typeafter	**-957.94**	**303.18**	**35.33**	**-3.153**	**< .01**
Typeif	180.27	247.59	72.91	0.726	.4703

Table C 2:: Summary of Linear Mixed Effects Model for response times: effects of AgeGroup and Type. The reference levels are for AgeGroup: 4years, for Type: because. Significant effects are highlighted in bold font.

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	5805.84	417.6	102.61	13.903	< $2e^{-16}$
AgeGroup5	-1750.53	417.22	65.68	-4.196	$8.34E^{-05}$
Typeafter	**-1227.54**	**365.56**	**56.32**	**-3.358**	**< .01**
Typebefore	**-1138.22**	**337.53**	**46.96**	**-3.372**	**<.01**
Typebecause	-180.18	247.81	63.97	-0.727	.46982

Table C 3: Summary of Linear Mixed Effects Model for response times: effects of AgeGroup and Type. The reference levels are for AgeGroup: 4years, for Type: if. Significant effects are highlighted in bold font.

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	877.5	190.9	11.4	4.596	< 0.0017
Typeafter	**206.4**	**103.5**	**433.7**	**1.995**	**< 0.05**
Typebecause	207.4	106.2	53.5	1.954	.056
Typeif	**271.9**	**108.1**	**54.4**	**2.514**	**< 0.05**

Table C 4: Summary of Linear Mixed Effects Model for response times for the adult group: effect of Type. The reference level is “after”. Significant effects are highlighted in bold font. Note that because Type has four levels, this table shows the results only for the model with “before” as the reference level.

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	1084.92	191.06	11.4	5.678	.000123
Typeafter	-0.99	106.65	54.3	-0.009	.992627
Typebefore	-207.42	106.17	53.5	-1.954	.055967
Typeif	64.44	105.54	464.7	0.611	.541794

Table C 5: Summary of Linear Mixed Effects Model for response times for the adult group: effect of Type. The reference level is “after”. Significant effects are highlighted in bold font. Note that because Type has four levels, this table shows the results only for the model with “because” as the reference level.

	Estimate	Std. Error	df	t value	Pr(>ltl)
(Intercept)	1149.35	192.17	11.7	5.981	$7.10E^{-05}$
Typeafter	-65.43	108.56	55	-0.603	.5492
Typebefore	**-271.86**	**108.14**	**54.4**	**-2.514**	**< 0.05**
Typebecause	-64.44	105.54	408.4	-0.611	.5418

Table C 6: Summary of Linear Mixed Effects Model for response times for the adult group: effect of Type. The reference level is “after”. Significant effects are highlighted in bold font. Note that because Type has four levels, this table shows the results only for the model with “if” as the reference level.

Appendix D

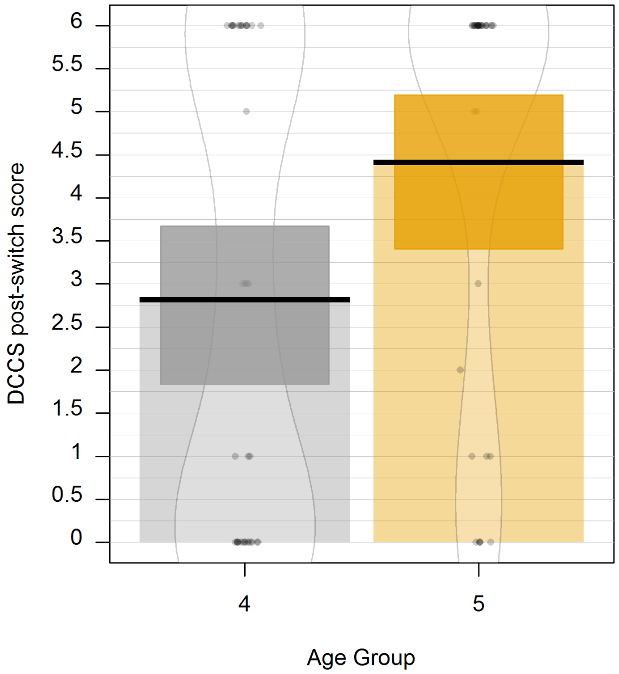


Figure D 1: Individual dots represent individual scores (raw data). Bars indicate means, beans indicate smoothed density, and bands indicate the 95% Bayesian Highest Density Interval (HDI). The pirate plot has been produced using the R package "yarrr" (Phillips, 2016).

	Standard correlation				**Bayesian correlation**	
Task	*r*	*t*	*df*	*p*	*r*	*BF*
CELF Linguistic Concepts	0.43	3.96	70	< .001*	0.41	102.4$^{\diamond}$
CELF Sentence Structure	0.32	2.85	70	< .01*	0.31	4.16$^{\diamond}$
ERB PSRep	0.24	2.03	69	< .05*	0.25	0.68
ERB Sentence Imitation	0.38	3.38	69	< .01*	0.37	15.56$^{\diamond}$
Day/Night	0.02	0.15	70	.87	0.02	0.09$^{\diamond}$
DCCS post-switch	0.31	2.71	70	< .01*	0.30	2.9

Table D 1: Correlation coefficients, t-values, degrees of freedom (df), probabilities (p), correlation coefficients obtained through Bayesian tests, and Bayes factors (BF) for the correlations between standardised test scores (z-scores) and mean accuracy. Asterisks indicate statistical significance; diamonds indicate at least substantial evidence (for the H_0, that is, no correlation if below 1/3, for the $H_{A,}$ if above 3).

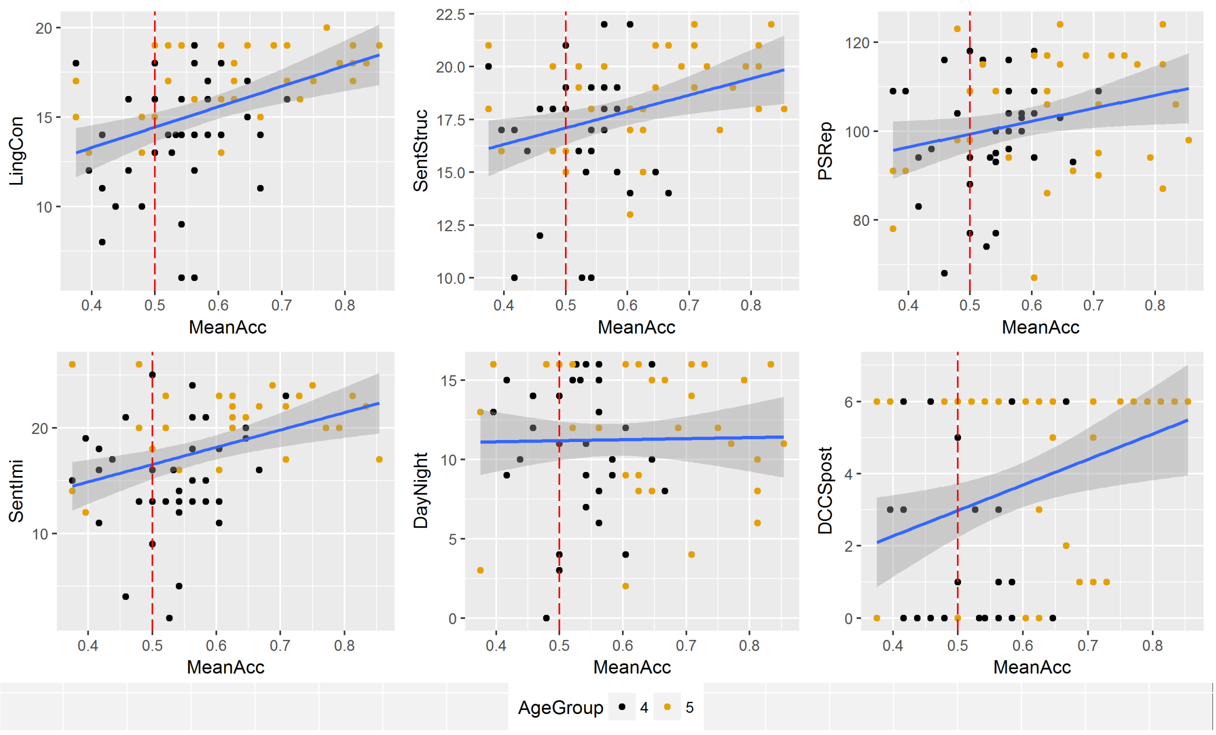


Figure D 2: Scatterplots showing the relationship between mean accuracy (MeanAcc) and the raw test scores for both age groups in the Linguistic Concepts sub-test, the CELF Sentence Structure sub-test, the ERB Preschool Repetition subtest (PSRep), the ERB Sentence imitation sub-test, the Day/Night task, and the DCCS post-switch phase. Blue lines indicate smoothed conditional means, grey shades indicate confidence intervals. Red dotted lines indicate chance level.

	Standard correlation				Bayesian correlation	
Task	*r*	*t*	*df*	*p*	*r*	*BF*
CELF Linguistic Concepts	-0.37	-3.30	69	< .01*	-0.36	13.79$^{\lozenge}$
CELF Sentence Structure	-0.24	-2.05	69	< .05*	-0.23	0.70
ERB PSRep	-0.06	-0.5	68	.6194	-0.06	0.11
ERB Sentence Imitation	-0.22	-1.88	68	.06	-0.22	0.51
Day/Night	-0.10	-0.87	69	.39	-0.10	0.13
DCCS post-switch	-0.26	-2.25	69	< .05*	-0.25	1.04

Table D 2: Correlation coefficients, t-values, degrees of freedom (df), probabilities (p), correlation coefficients obtained through Bayesian tests, and Bayes factors (BF) for the correlations between standardised test scores (z-scores) and mean response times. Asterisks indicate statistical significance; diamonds indicate at least substantial (for the H_0, that is, no correlation if below 1/3, for the $H_{A,}$ if above 3).

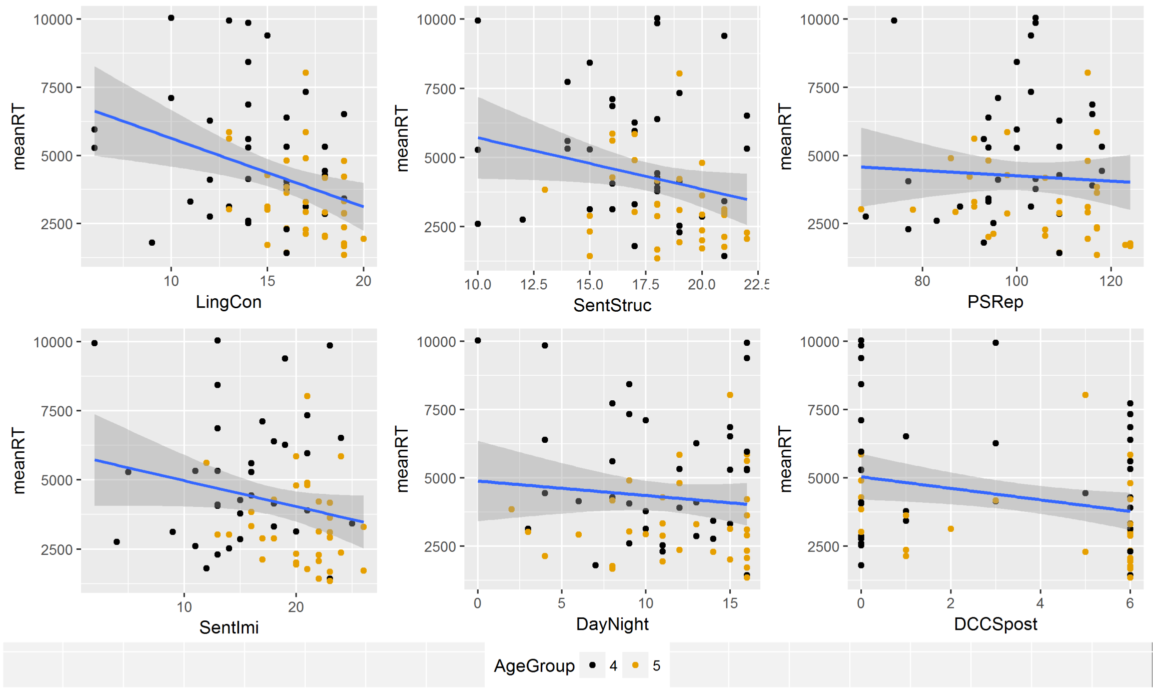


Figure D 3:Scatterplots showing the relationship between mean response time (MeanRT) and the raw test scores for both age groups in the Linguistic Concepts sub-test, the CELF Sentence Structure sub-test, the ERB Preschool Repetition subtest (PSRep), the ERB Sentence imitation sub-test, the Day/Night task, and the DCCS post-switch phase. Blue lines indicate smoothed conditional means, grey shades indicate confidence intervals.

	Estimate	Std. Error	z value	Pr(>lzl)
(Intercept)	0.1	0.2	0.402	.69
AgeGroup5	**1.1**	**0.2**	**4.457**	**< .0001**
Typeafter	0.1	0.2	0.383	.70
Typebecause	0.3	0.2	1.383	.17
Typeif	0.2	0.2	0.966	.33
ClauseOrdersub-main	0.2	0.2	1.213	.23
scale(LingCon)	**0.2**	**0.1**	**2.446**	**< .05**
AgeGroup5:Typeafter	**-1.4**	**0.3**	**-4.701**	**< .0001**
AgeGroup5:Typebecause	**-1.2**	**0.3**	**-4.146**	**< .0001**
AgeGroup5:Typeif	**-1.5**	**0.3**	**-5.002**	**< .0001**
AgeGroup5:ClauseOrdersub-main	**-0.7**	**0.3**	**-2.395**	**< .05**
AgeGroup4:Typeafter:ClauseOrdersub-main	-0.2	0.3	-0.755	.45
AgeGroup5:Typeafter:ClauseOrdersub-main	**1.4**	**0.3**	**4.389**	**< .0001**
AgeGroup4:Typebecause:ClauseOrdersub-main	-0.2	0.3	-0.79	.43
AgeGroup5:Typebecause:ClauseOrdersub-main	**1.3**	**0.3**	**4.016**	**< .001**
AgeGroup4:Typeif:ClauseOrdersub-main	-0.3	0.3	-1.173	0.24
AgeGroup5:Typeif:ClauseOrdersub-main	1.4	0.3	**4.636**	**< .0001**

Table D 3: Summary of Generalized Linear Mixed Effects Model for the log odds for accuracy responses: effects and Interactions of AgeGroup, Type, and ClauseOrder. The reference levels are for AgeGroup: 4years, for Type: before, and for ClauseOrder: main-subordinate. Significant effects are highlighted in bold font.

	Estimate	Std. Error	z value	Pr(>lzl)
(Intercept)	0.4	0.2	2.178	.48
AgeGroup5	-0.2	0.2	-0.707	.31
Typeafter	-0.2	0.2	-1.025	.17
Typebefore	-0.3	0.2	-1.378	66
Typeif	-0.1	0.2	-0.446	.92
ClauseOrdersub-main	**0.0**	**0.2**	**0.095**	**< .05**
scale(LingCon)	0.2	0.1	2.445	.57
AgeGroup5:Typeafter	-0.2	0.3	-0.565	**< .0001**
AgeGroup5:Typebefore	1.2	0.3	4.14	.38
AgeGroup5:Typeif	**-0.3**	**0.3**	**-0.886**	**< .01**
AgeGroup5:ClauseOrdersub-main	0.8	0.3	2.592	.97
AgeGroup4:Typeafter:ClauseOrdersub-main	0.0	0.3	0.035	.75
AgeGroup5:Typeafter:ClauseOrdersub-main	0.1	0.3	0.323	.43
AgeGroup4:Typebefore:ClauseOrdersub-main	**0.2**	**0.3**	**0.785**	**< .0001**
AgeGroup5:Typebefore:ClauseOrdersub-main	-1.3	0.3	-4.013	.70
AgeGroup4:Typeif:ClauseOrdersub-main	-0.1	0.3	-0.386	.55
AgeGroup5:Typeif:ClauseOrdersub-main	0.2	0.3	0.602	.48

Table D 4: Summary of Generalized Linear Mixed Effects Model for the log odds for accuracy responses: effects and Interactions of AgeGroup, Type, and ClauseOrder. The reference levels are for AgeGroup: 4years, for Type: before, and for ClauseOrder: main-subordinate. Significant effects are highlighted in bold font.

	Estimate	Std. Error	z value	Pr(>lzl)
(Intercept)	0.3	0.2	1.651	.10
AgeGroup5	-0.4	0.2	-1.83	.07
Typeafter	-0.1	0.2	-0.611	.54
Typebefore	-0.2	0.2	-0.967	.33
Typebecause	0.1	0.2	0.45	.65
ClauseOrdersub-main	-0.1	0.2	-0.448	.65
scale(LingCon)	**0.2**	**0.1**	**2.444**	**< .05**
AgeGroup5:Typeafter	0.1	0.3	0.325	.74
AgeGroup5:Typebefore	**1.5**	**0.3**	**5.002**	**< .0001**
AgeGroup5:Typebecause	0.3	0.3	0.882	.38
AgeGroup5:ClauseOrdersub-main	**1.0**	**0.3**	**3.629**	**< .001**
AgeGroup4:Typeafter:ClauseOrdersub-main	0.1	0.3	0.42	.67
AgeGroup5:Typeafter:ClauseOrdersub-main	-0.1	0.3	-0.288	.77
AgeGroup4:Typebefore:ClauseOrdersub-main	0.3	0.3	1.175	.24
AgeGroup5:Typebefore:ClauseOrdersub-main	**-1.4**	**0.3**	**-4.635**	**< .0001**
AgeGroup4:Typebecause:ClauseOrdersub-main	0.1	0.3	0.38	.70
AgeGroup5:Typebecause:ClauseOrdersub-main	-0.2	0.3	-0.599	.55

Table D 5: Summary of Generalized Linear Mixed Effects Model for the log odds for accuracy responses: effects and Interactions of AgeGroup, Type, and ClauseOrder. The reference levels are for AgeGroup: 4years, for Type: if, and for ClauseOrder: main-subordinate. Significant effects are highlighted in bold font.

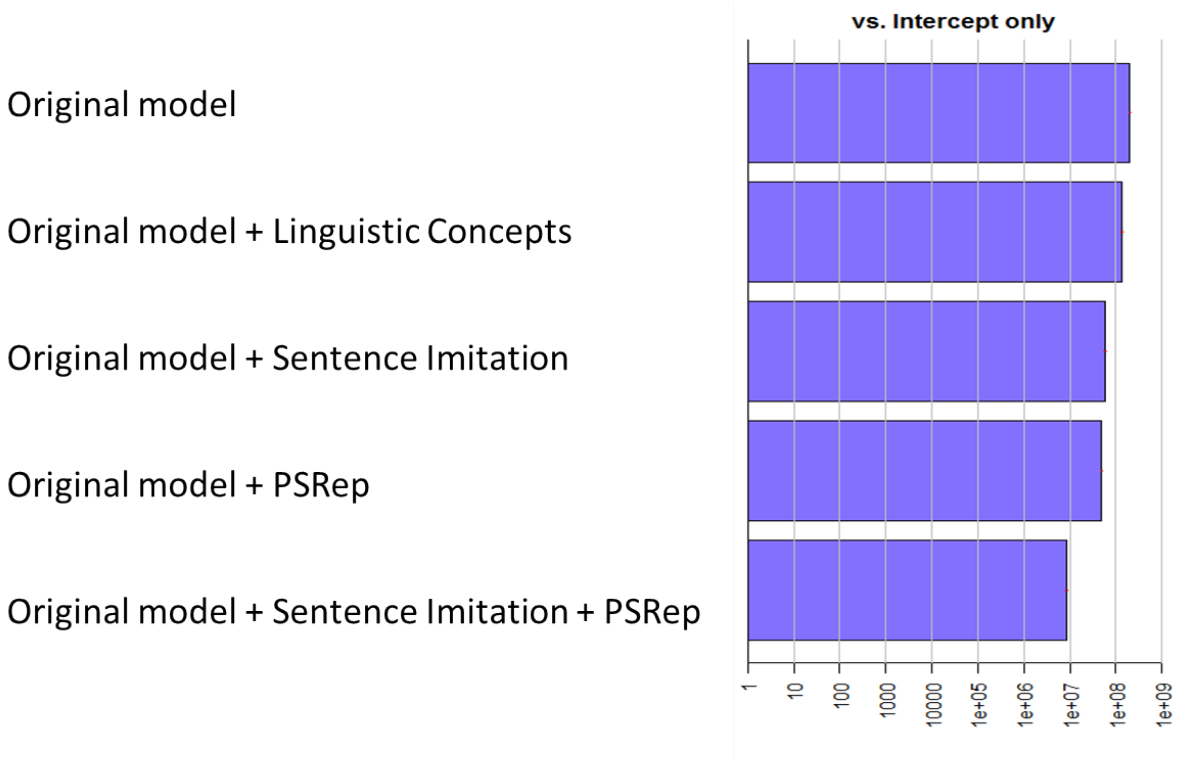


Figure D 4: Bayes factors for five different models predicting accuracy, compared to the null model (intercept).