View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by City Research Online

City Research Online

City, University of London Institutional Repository

Citation: Polisenska, K., Chiat, S. ORCID: 0000-0002-8981-8153, Szewczyk, J. and Twomey, K. E. (2021). Effects of semantic plausibility, syntactic complexity and n-gram frequency on children's sentence repetition. Journal of Child Language, 48(2), pp. 261-284. doi: 10.1017/S0305000920000306

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/26246/

Link to published version: https://doi.org/10.1017/S0305000920000306

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

City Research Online

City, University of London Institutional Repository

Citation: Polisenska, Kamila, Chiat, Shula ORCID: 0000-0002-8981-8153, Szewczyk, Jakub and Twomey, Katherine E (2021). Effects of semantic plausibility, syntactic complexity and n-gram frequency on children's sentence repetition. JOURNAL OF CHILD LANGUAGE, 48(2), doi: 10.1017/S0305000920000306

This is the draft version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/26246/

Link to published version: http://dx.doi.org/10.1017/S0305000920000306

Copyright and reuse: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Effects of semantic plausibility, syntactic complexity and n-gram frequency on children's sentence repetition

Kamila POLIŠENSKÁ

The University of Manchester, UK

Shula CHIAT

City, University of London, UK

Jakub SZEWCZYK

Institute of Psychology, Jagiellonian University, Kraków, Poland

Katherine E. TWOMEY

The University of Manchester, UK

Corresponding author:

Kamila Polišenská

Division of Human Communication, Development and Hearing

The University of Manchester

Oxford Road, Manchester M13 9PL, United Kingdom

Phone: +44 0161 275 3369

Email: kamila.polisenska@manchester.ac.uk

Keywords: sentence repetition, n-grams, semantics, syntax

Abstract

Theories of language processing differ with respect to the role of abstract syntax and semantics vs surface-level lexical co-occurrence (n-gram) frequency. The contribution of each of these factors has been demonstrated in previous studies of children and adults, but none have investigated them jointly. This study evaluated the role of all three factors in a sentence repetition task performed by children aged 4-7 and 11-12 years. It was found that semantic plausibility benefitted performance in both age groups; syntactic complexity disadvantaged the younger group but benefitted the older group; while contrary to previous findings, n-gram frequency did not facilitate, and in a post hoc analysis even hampered, performance. This new evidence suggests that n-gram frequency effects might be restricted to the highly constrained and frequent n-grams used in previous investigations, and that semantics and morphosyntax play a more powerful role than n-gram frequency, supporting the role of abstract linguistic knowledge in children's sentence processing.

Introduction

Theories of language processing vary widely in the role they attribute to item-specific versus abstract knowledge, from almost-total reliance on storage and reuse of exemplars (e.g. Ambridge, 2019; Bybee, 2001, 2009) to almost-total reliance on generative linguistic capacity (e.g. Hauser, Chomsky, & Fitch, 2002), with many in between (e.g. Goldberg, 2003). The definition of grammar and lexicon and the distinction between these will depend on where on the spectrum the theories lie. On the one hand, language has been seen as driven by rule-based decomposition of hierarchical syntactic structure, with a clear distinction between grammar and lexicon. In contrast, a growing number of theories do not draw a sharp distinction between the lexicon and the combinatory rules (e.g. Langacker, 1987; Bybee, 2001, 2006; Goldberg, 2003). These usage-based approaches claim that people mentally store exemplars, or tokens of linguistic experience, which can be larger than single words. From these exemplars, speakers generalize at multiple levels of granularity (e.g. morpheme, word, or phrase), and these generalizations give rise to grammatical knowledge. As a result, single words and multi-word chunks are processed and stored similarly. Most theories now accept the importance of input statistics to various degrees (Ambridge, Kidd, Rowland, & Theakston, 2015), but the debate continues about the relative importance of stored items vs abstract syntactic structure in language use.

Between the most polarized ‘generative’ and ‘usage-based’ accounts are theories highlighting the confluence of multiple levels of knowledge – from lexically-specific multi-word chunks through morphosyntactic configurations to fully abstract syntactic structures that are independent of lexical content – corresponding to multiple levels of abstraction (e.g., Chiat, 2001; Christiansen & Chater, 2016; Culicover & Jackendoff, 2005). According to such theories, multiple levels of knowledge are always available, but their contribution may vary depending on the language behavior/task and content (Frank & Christiansen, 2018). Their contribution may also vary according to age (Abbot-Smith & Behrens, 2010; Tomasello, 1992).

The current study addresses this ‘multi-level’ view of language knowledge in the particular case of immediate sentence repetition. The aim was to evaluate the contribution of different levels of

knowledge at two ages: when production of complex sentences is still emerging, and when production of complex sentences is well established.

Immediate sentence repetition

Immediate sentence repetition (SR) has been used to investigate language processing in children and adults (e.g. Bannard & Matthews, 2008; Kidd, Brandt, Lieven, & Tomasello, 2007; Valian, Prasada, & Scarpa, 2006). While SR tasks present participants with a string of phonological input and simply ask them to reproduce this, it is clear that respondents are not simply turning a string of phonological input into a string of phonological output. Rather, processing and production in this task are affected by a complex combination of factors. Most obviously, recall capacity is substantially greater for strings of words organized into sentences than randomly ordered (e.g., Jefferies, Ralph, & Baddeley, 2004; Polišenská, Chiat, & Roy, 2015), suggesting that syntactic and semantic knowledge affects processing. Accordingly, SR is taken to be a language-sensitive measure evaluating language processing and knowledge, rather than a purely quantitative measure of verbal short-term memory (e.g. Klem, Melby-Lervåg, Hagtvet, Lyster, Gustafsson, & Hulme, 2015; Polišenská et al., 2015; Riches, 2012).

Most research on children's SR has isolated the influence of syntactic factors (e.g., structural complexity: Frizelle & Fletcher, 2014; Riches, 2012; Kidd et al., 2007; Moll, Hulme, Nag, & Snowling, 2015), semantic factors (e.g., plausibility: Miller & Isard, 1963; Polišenská et al., 2015; Valian et al., 2006; Wallan, Chiat, & Roy, 2011) and/or one quantitative factor (e.g., sentence length: Moll et al., 2015), but has not considered the role of these factors together. Both semantic plausibility and syntactic complexity involve higher level linguistic knowledge, but the influence of both of these factors could be a product of purely surface-level distributional knowledge, i.e. familiarity with sequences of words, which does not require knowledge of words' meanings or referents in the real world (semantic knowledge) or knowledge of structural relations and operations (e.g. movement, embedding). Participants might be better at repeating sentences containing frequent sequences of words due to implicit, statistical learning and hence show better recognition, retention, and production of these sequences (Dell & Chang, 2014). For example, plausible sentence fragments might occur much more frequently than implausible ones, e.g. *red apple* vs. *stripy apple*. To what extent, then,

might surface distributional frequency account for effects of semantic and syntactic factors? In the following sections we review empirical evidence for the influence of syntactic knowledge, semantic information and frequency on SR, before considering theoretical approaches that predict the influence of all three and raising questions about their relative contribution.

Syntactic knowledge: the effect of complexity

A number of studies have established that children's recall is poorer for syntactically complex than for syntactically simple sentences (Frizelle & Fletcher, 2014; Riches, 2012; Kidd et al., 2007; Moll et al., 2015). In a study seeking to disentangle the contributions of memory capacity and language in SR, Moll and colleagues (2015) manipulated length and syntactic complexity. They administered their SR task to groups of children with and without dyslexia, and found that length and syntactic complexity each had significant and independent effects. These findings demonstrate that although complex sentences are typically longer than simple sentences, effects of syntactic complexity on SR were not purely due to length and hence verbal memory capacity. Effects of syntactic complexity have also been examined in studies investigating repetition of relative clause structures in children with and without language impairment (Diessel & Tomasello, 2005; Riches, 2017). Findings revealed a hierarchy of difficulty: children performed better on subject than object relatives, with indirect object relatives, oblique object relatives and genitive relatives yielding progressively poorer performance. Thus, a sentence's syntactic complexity appears to affect its accessibility, suggesting that structural knowledge plays a role in sentence repetition. Given that children's acquisition of complex structures is gradual, we might expect effects of syntactic complexity on SR to change with age and development.

However, syntax may not be the only or even key factor in this observed pattern of performance. In a study of SR performance in both English and German, Kidd et al. (2007) showed that the difference between repetition of subject and object relative clauses disappeared when the object relative contained an inanimate head noun and a pronominal subject (*inanimate NP – that – pro* as in 'There is [the book] [that you read…]'). This finding was replicated by Frizelle and Fletcher (2014) in children with Specific Language Impairment (now known as Developmental Language Disorder) as well as age-matched and younger typically-developing controls. Kidd and colleagues

point out that children's preference for this semantic-morphosyntactic configuration reflects distributional patterns in the input they receive (i.e., distributional frequencies of complex structures). Collectively, then, these studies provide evidence that syntactic complexity and the internal morphosyntactic structure of complex sentences affect SR independently of sentence length, but leave open the possibility that these effects arise from differences in semantic information and/or input frequency.

Semantic knowledge: the effect of plausibility

Several studies have revealed that participants recall semantically plausible sentences better than implausible/anomalous sentences. Evidence is available for both children and adults and for typologically different languages (English, Arabic, Czech; e.g. Miller & Isard, 1963; Polišenská et al., 2015; Valian et al., 2006; Wallan et al., 2011). For example, in a delayed recall task, Polišenská, Chiat, Comer and McKenzie (2014) demonstrated that plausibility improved recall in adults and six-year-old children, and that this effect increased with task difficulty. Plausible and implausible sentences encode different situations, with plausible sentences describing situations more likely to have occurred in the real world and therefore to have been experienced and encoded by the listener. As a result, plausible sentences may activate stronger semantic representations, facilitating recall. Again, however, these studies do not take into account the possibility that sequences of words in plausible sentences are more frequent in the input.

Knowledge of lexical sequences: the effect of n-gram frequency

A large body of research demonstrates that input frequency affects language processing across development (for a review, see Ambridge et al., 2015). In particular, recent empirical and computational research has suggested a role for the chunk – or n-gram – in language processing. N-grams are sequences of two, three or more words and are not limited to constituents, clauses or intonational phrases. Any sentence can be broken into n-grams of different levels of granularity. Consider the sentence '*The bird picked a red apple from the garden*'. This sentence has ten 2-grams, e.g. '*the bird'*, '*picked a*', '*a red*', nine 3-grams, e.g. '*a red apple*' '*red apple from'*, eight 4-grams, e.g. '*bird picked a red*', '*apple from the garden*', and seven 5-grams, e.g. '*the bird picked a red*', '*picked a red apple from*'. As is clear from the examples, some n-grams also form a constituent, e.g.

'*a red apple*' while others cross syntactic boundaries, e.g. '*picked a*'. If such sequences of words are stored in long-term memory in addition to individual words, it follows that this knowledge of chunks might contribute to SR, with sequences containing more frequent n-grams producing better recall.

A growing literature on a variety of tasks has shown that adult processing of multi-word sequences is affected by n-gram frequency. Arnon and Snider (2010) found that high-frequency multi-word phrases were processed faster by adults than strings of lower frequency in a series of phrasal decision tasks. A recent study by Supasiraprapa (2019) replicated this finding using the same stimuli with native and non-native speakers of English. Using a reading task, Tremblay, Derwing, Libben and Westbury (2011) showed that this was even true when n-grams straddled syntactic constituents, with non-constituent sequences of higher n-gram frequency read faster than their lower frequency counterparts (e.g. '*in the middle of the*' vs. '*in the front of the*'). Similarly, Arnon and Priva (2013) found effects of multi-word frequency on phonetic duration in elicited and spontaneous production of word sequences both within and across syntactic boundaries. Turning to children, effects of n-gram frequency have been found in just one experimental study with 2- and 3-year-olds: Bannard and Matthews (2008) compared the children's repetition of four-word targets that were identical apart from the final word (e.g. '*a drink of milk*' vs. '*a drink of tea*'), and demonstrated that they were significantly better at repeating the higher-frequency chunks in terms of both accuracy and speed of recall. This empirical work is supported by computational simulations in which n-gram-based learning mechanisms produce developmental trajectories which reflect those seen in child language production (e.g. McCauley & Christiansen, 2017). Again, however, these studies do not address the roles of abstract knowledge and semantics in processing.

Aims and hypotheses

Previous research on immediate sentence repetition has for the most part focused on one level of language knowledge. Our study aimed to shed new light on current debates by investigating the contribution of several levels of language knowledge simultaneously, and whether this contribution changes between the early primary school and early secondary school years. In particular, we examine the contribution of frequency of words and multi-word sequences (n-gram frequency), syntactic complexity, and semantic plausibility. Including n-gram frequency of targets in our analyses allows us

to explore the independent contribution of surface-level knowledge of word combinations. If plausibility or complexity contribute independently of n-gram frequency, this would suggest that children use sources of information in SR that go beyond frequency of word co-occurrence and rely on some level of abstraction, therefore findings would point towards theories that stress the role of frequency and abstract knowledge. However, it is also possible that n-gram frequency is of primary importance in children's recall performance. In this case, previously reported findings on the effect of plausibility and/or complexity in recall could instead be due to n-gram frequency and better explained by frequency only theoretical accounts. This possibility is particularly important to our interpretation of existing findings, because what might be traditionally attributed to high-level processing (syntactic knowledge, semantic knowledge) might, to a greater or lesser extent, be explained on a much lower level (n-gram frequency).

We hypothesize that higher n-gram frequency, simple syntactic structure and greater semantic plausibility will benefit performance in both groups of children. We further hypothesize that, while older children's repetition will be more accurate due to increased knowledge at all levels, the relative contribution of different levels will change. In particular, we hypothesize that older children's repetition of complex sentences will benefit from abstract distributional knowledge that is less available or robust in younger children. This is based on the theory that knowledge of complex sentence structure arises through abstraction across multi-word chunks, and that complex structures are still emergent in the early primary years but adult-like by the early secondary years (Ambridge, Barak, Wonnacott, Bannard, & Sala, 2018; Bowerman, 1988). In contrast, while older children have greater experience and understanding of real world situations which contributes to their growing knowledge of semantic-morphosyntactic configurations, we would not expect this knowledge to benefit sentences that contravene real world experience and semantic knowledge. Hence, we expect implausibility to be similarly detrimental for repetition of sentences regardless of age.

Previous studies investigating the effects of n-gram frequency were carried out with adults, with the exception of Bannard and Matthews (2008), which focused on 2- to 3-year-old children in the early stages of language acquisition. Participants in the current study were primary and secondary school-aged children. This is the first study to examine n-gram effects across this age range and our

expectations regarding interaction between n-gram frequency and age are less clear. In Bannard and Matthews' study, the 3-year-old children performed better than 2-year-olds but there was no significant interaction between age and frequency. The authors interpreted these findings in favour of continuity in frequency effects across development, while also pointing out the possibility that the frequency effect might diminish in older children and adults. We know older children have greater experience and knowledge of words and multi-word chunks, and we expect relatively high frequency items to elicit more accurate performance than relatively low frequency items regardless of age. However, this does not take account of the possibility of substantial changes in the relative input frequency of specific n-gram targets across the age range (see Discussion).

Based on these hypotheses, we predict independent contributions of age, multi-word (n-gram) frequency, syntactic complexity, and plausibility to children's immediate sentence repetition. In addition, we predict an interaction between age and complexity, with the effect of complexity reducing across age.

Method

Participants. Fifty children across two age groups participated in the study: a younger group (n = 21, 10 male, mean age = 73 months, *SD* = 11 months, age range 56 - 90 months) and older group (n = 29, 14 male, mean age = 148 months, *SD* = 5 months, age range 135 - 155 months). Children were recruited via schools in the South and East of England. Inclusion was based on meeting all of the following criteria: English as a first language, no concerns about children's development and their parents gave written consent. Ethical approval for this study was granted by the City University School of Health Sciences Ethics Committee.

Materials

Sentence repetition task. The targets comprised 36 sentences (see Appendix), spanning a spectrum of plausibility, from semantically plausible to semantically implausible (see plausibility ratings below). Sentences further differed in syntactic complexity (simple vs complex), with complex sentences containing two clauses while simple sentences contained only one clause. All target sentences comprised nine words with five content and four function words. In addition, all sentences

contained 11 syllables. Semantically, sentences were all drawn from the same range of verbs (action, location, possession, state, mental state) and nouns (all chosen to be familiar to the participants). Function words could occur more than once, reflecting their high token frequency in English sentences (Shi, Morgan, & Allopenna, 1998).

Semantic plausibility. Semantically implausible (SI) sentences were created by replacing the content words in semantically plausible (SP) sentences with different content words that were matched for number of syllables, and chosen to violate selectional restrictions (Chomsky, 1965; Hare, McRae, & Elman, 2003) resulting in semantically anomalous sentences such as the following:

SP: The	man	that	the	girl	saw	posted	a	letter
SI: The	boat	that	the	bird	wore	folded	a	penny

This process ensured that SP and SI targets were matched for number of syllables and number of words in addition to syntactic and prosodic structure.

In order to establish the perceived plausibility of these sentences, we presented them to 23 adult native English speakers in the same order as presented for the sentence repetition task. After hearing the full set once, participants were instructed to listen to each sentence in turn, and rate the plausibility of the sentence using a 7-point scale ranging from ‘completely meaningless’ to ‘completely meaningful’, beginning with four practice items. The task was presented and responses recorded using Qualtrics software (©2019, Qualtrics®, Provo, UT). It was emphasised that the ‘meaningfulness’ judgement referred to how much the sentence made sense, rather than the likelihood of the situation it was describing, i.e. sentences that are considered ‘plausible’ have a clear, sensible meaning which is easy to interpret, whereas for implausible sentences the meaning is bizarre and harder to interpret. The mean plausibility rating was 4.1 (*SD* = 2.5) and although the distribution was bimodal, there was some variance within plausible and implausible sentences. In the analyses, we used the mean plausibility rating for each item as a continuous variable.

Syntactic complexity. Two levels of complexity were included:

Simple: Targets in the simple condition were simple sentences with one clause only and included a range of structures: active sentences, passive sentences, two double-object dative structures, possessives embedded in a NP.

Complex: Targets in the complex condition contained embedded sentences in the form of relative clauses, which are particularly complex structures for children to process (Kidd & Bavin, 2002). The relative clauses were either subject-object relatives, as in '*the train that the boy missed was very busy'*; object-object, as in '*the queen hated the dress that the lady bought*'; or subject-possessive, as in '*the driver whose coat was dirty sold the car*'. The relative pronoun *that* was always included even if it was optional. Semantically, sentences with relative clauses contained an event/predication within an event/predication.

N-gram frequency. To obtain the word n-gram frequency index, we extracted from each sentence all possible n-grams (with *n* in the range 2 - 5): eight 2-grams, seven 3-grams, six 4-grams and five 5-grams, totalling 26 n-grams per sentence. We then checked the frequency of each n-gram in the Google Books Ngram Corpus[1], containing over 4% of books ever printed (version 2; Michel et al., 2011). Next, for each word in each sentence we computed the mean log n-gram frequencies (separately for n = 2, 3, 4, 5; logarithms were taken before computing the mean) for all n-grams in which the word occurred. For example, in the sentence *The man that the girl saw posted a letter*, the verb *saw* occurs in two 2-grams (*girl saw*, *saw posted*), three 3-grams (*the girl saw*, *girl saw posted*, *saw posted a*) and so on. In this way, for each data-point (a word in a sentence) we had four mean log n-gram-frequencies. Because the indices of frequency for n-grams of different length are highly correlated (see Table 1), and we had no *a priori* hypotheses related to the contribution of n-grams of

[1] Google Books N-gram Corpus, which was the source of n-gram frequencies in our study, is not a direct measure of the input that children receive. In contrast, studies of the effects of construction frequency have used child-directed speech as the source (Bannard & Matthews, 2008; Kidd et al., 2007). The children in Bannard and Matthews (2008) were younger (2- and 3-year-olds) so the frequencies based on child-directed speech were more important than in our study where the children were older and therefore more likely to be exposed to written as well as spoken input, and the sentence targets were unlike input to 2- to 3-year-olds. There were two main advantages of using Google Books N-gram corpus: i) it is the largest database available, ii) it reflects the collective contribution of thousands of speakers rather than the speech of one individual (i.e. one child's mother in Bannard & Matthews, 2008).

different length, we extracted a common factor reflecting n-gram frequency using principal component analysis and used it in the analyses. For the main analysis, we extracted the first component only, but later in exploratory analyses we checked if the second component should also be included.

This n-gram component is a measure of how much support a word in a sentence may receive from various chunks of words in which the word occurs within a given sentence provided the child knows these chunks of words. Finding that repetition accuracy for the words in a sentence is co-predicted by this measure of words' n-gram frequency would indicate that the child knows those chunks of words and that they help in encoding and/or maintaining the word, as part of the sentence, in the SR task. For a similar logic for phoneme-based n-grams, see Szewczyk, Marecka, Chiat and Wodniecka (2018).

Table 1. *Correlation matrix between Complexity, Plausibility, Lexical Frequency, and 2-, 3-, 4-, and 5-gram log frequencies*

	Complexity	Plausibility	Lexical freq.	2-gram freq.	3-gram freq.	4-gram freq.
Plausibility	-0.05					
Lexical freq.	-0.03	0.04				
2-gram freq.	-0.12	0.15	0.69			
3-gram freq.	-0.17	0.34	0.32	0.69		
4-gram freq.	-0.12	0.33	0.26	0.41	0.62	
5-gram freq.	-0.09	0.24	0.09	0.21	0.35	0.61

Procedure. Children were seen individually in a quiet room in a single testing session. The sentences were recorded by a female native English speaker with a southern British accent and the order of the test sentences was randomized and presented on a laptop through children's headphones. The presentation started with two practice sentences of six words each to familiarize children with the repetition task. Administration of the task was controlled by the researcher. Each stimulus was only played once and a child was asked to repeat the sentence played. If a child did not respond for ten seconds after the presentation of any sentence, the researcher proceeded to the next sentence.

Recordings were not replayed even if requested. Children were praised regardless of accuracy, and no feedback was given as to whether the child's response was correct or incorrect. Participants' responses were recorded on a Marantz Professional PMD620 digital recorder. At the end of the session, children were praised for having completed the task and thanked for their participation.

Scoring and inter-rater reliability. Each sentence was scored for the total number of words repeated correctly (maximum nine per sentence). Words were considered to be correct regardless of their position in the sentence. Omissions of inflections and over-regularizations were allowed. Twenty participants (40%) were independently transcribed and scored by a rater who was blind to the original scoring. The level of inter-rater reliability was assessed by intra-class correlation coefficient (ICC) and achieved the following level: ICC=0.94, CI: 0.93-0.95, which confirmed that the measure was highly reliable.

Statistical analyses. Data were analyzed using generalized mixed effects models, using the lme4 package for R (version 1.1-21; Bates, Mächler, Bolker, & Walker, 2015). The analyses were conducted on the level of individual words building the target sentence. The dependent variable coded whether a given word, which should have occurred in children's repeated sentence, did occur there, irrespective of its position in the repeated sentence. Because of the binomial nature of the dependent variable, we used the binomial distribution with logit link function. For visualization, we transformed data into probability of repeating a word.

In all analyses categorical variables were deviation coded, while continuous variables were centered, and thus the intercept reflects the mean value of the dependent variable across all conditions. Syntactic Complexity was coded to reflect the difference between complex and simple sentences. Figures show partial effects computed from model parameters, with all random variance removed.

We considered one subject-related predictor: Age; and several sentence-related predictors: Plausibility (continuous plausibility rating), Syntactic Complexity (simple, complex), (log) N-gram Frequency, Lexical Frequency (Zipf score from the Subtlex-UK corpus; van Heuven et al., 2014), and Word Position (position of the word in the target sentence). We also considered interactions of N-gram Frequency, Plausibility, and Syntactic Complexity with Age. We were not interested in the

effect of Word Position, but we assumed that it might play a role due to primacy effects, and thus controlling it may increase the estimation precision of other predictors.

For all analyses we use the maximum random effects structure with correlations between random effects removed. The random variables included participant and sentence (N-gram Frequency, Lexical Frequency and Word Position were within-sentence variables, Syntactic Complexity and Plausibility were between-sentence variables).

Results

To extract a common factor reflecting n-gram frequency, we ran a principal component analysis. Table 2 shows the resulting component loadings. As can be seen, the first principal component, accounting for 62% of variance, is loaded with the same sign by n-grams of all lengths, thus reflecting their common variance.

Table 2. Results of the principal component analysis on the 4 log n-gram frequency scores

	PC1	PC2	PC3	PC4
		Component Loadings		
mean 2-gram log frequency	-0.47	-0.58	0.51	-0.42
mean 3-gram log frequency	-0.55	-0.32	-0.29	0.71
mean 4-gram log frequency	-0.54	0.30	-0.59	-0.52
mean 5-gram log frequency	-0.43	0.68	0.56	0.21
		Variance Explained		
SD	1.57	0.96	0.61	0.49
Proportion of Variance	0.62	0.23	0.09	0.06
Cumulative Proportion	0.62	0.85	0.94	1.00

Next, we fitted the generalized mixed effects model (GLME), using the first principal component as the estimate of log frequency of all n-grams in which a given word occurred. The estimates of fixed and random effects are shown in Table 3.

Table 3. *Estimates of fixed and random effects in the GLME model*

Effect	Estimate	95% CI	*p*	by-Sentence SD	by-Subject SD
Intercept	2.93	2.52 – 3.33	< .0001	0.44	1.27
Complexity	0.04	-0.36 – 0.45	= .84	-	0.64
Plausibility	0.23	0.15 – 0.31	< .0001	-	0.12
Log N-gram Frequency PC1	-0.11	-0.25 – 0.03	= .11	0.31	0.13
Log Word Position	-0.38	-0.55 – -0.20	< .0001	0.39	0
Log Lexical Frequency	0.20	0.09 – 0.30	< .001	0.21	0.17
Age	0.56	-0.18 – 1.30	= .14	0.3	-
Complexity:Age	1.03	0.51 – 1.56	< .001	-	-
Plausibility:Age	0.03	-0.08 – 0.13	= .61	-	-
Log N-gram Frequency PC1:Age	-0.10	-0.26 – 0.06	= .21	0.24	-

The model reveals that on average, children repeated 95% of the target words (log-odds = 2.93). Words coming from more plausible sentences, as well as more frequent words, were repeated considerably better. Similarly, words occurring earlier in the sentence were more likely to be repeated. Across all children, Syntactic Complexity did not have a significant effect, but it interacted with Age: In the younger group, words from syntactically complex sentences were less likely to be repeated correctly than words from simple sentences. In the older group this effect reversed, with words from simpler sentences less likely to be repeated correctly. Finally, the first principal component of N-gram Frequency did not significantly predict accuracy of word repetition, nor did it enter into a significant interaction with Age. As a post-hoc analysis, we checked if the second principal component of N-gram Frequency improved the model's predictions. Model comparison showed that adding the component did not improve the model ($\chi^2(3) = 2.94$, $p = .4$).

Figure 1 shows the estimated predictions of each fixed effect (lexical frequency, word position, n-gram frequency, plausibility, and age x complexity interaction) based on the model shown in Table 3.

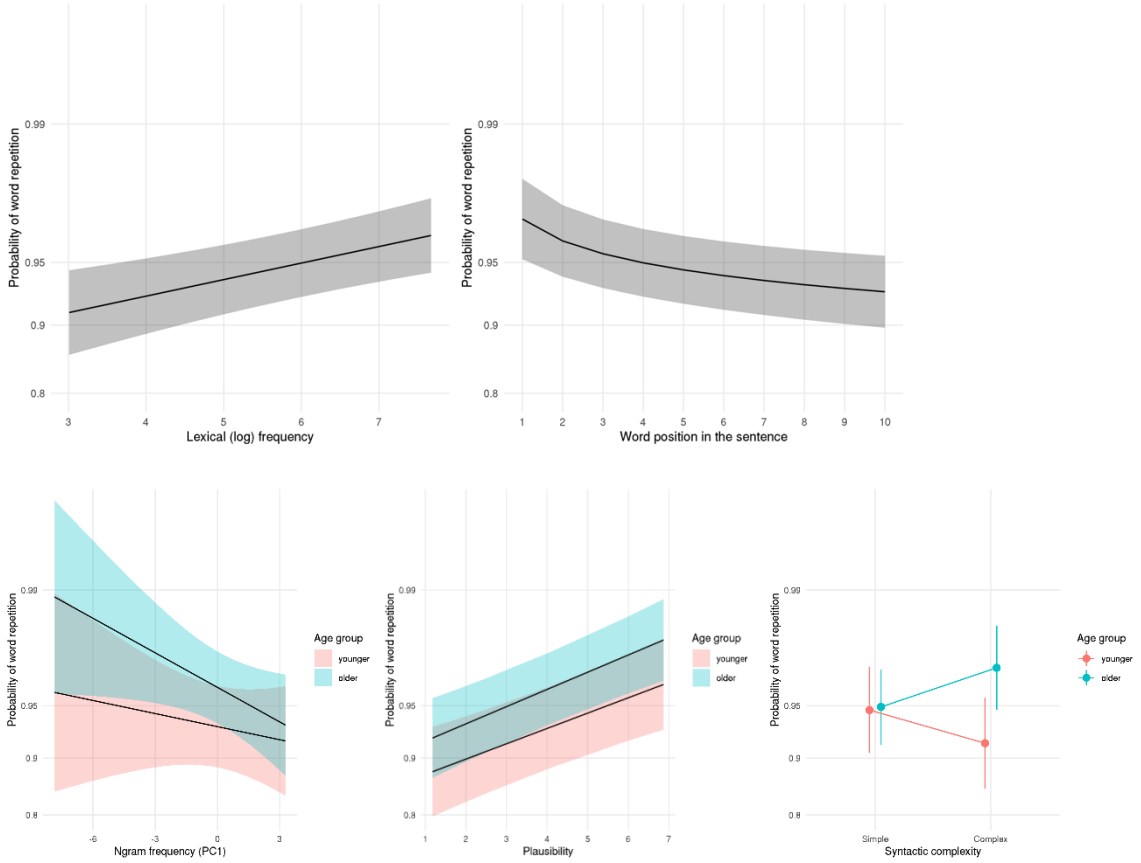

Figure 1. The estimated predictions of each fixed effect, based on the model shown in Table 3. Whiskers and bands show the 95% confidence interval with random variance excluded.

Post-hoc exploratory analysis of additional word-level effects. In the course of the review process we were asked to reanalyse the data with the inclusion of a third random variable – word. This modification in the analysis very significantly improved the model's fit to the data, and led to a notable change in the results. While the previously-found effects of plausibility and complexity were preserved, effects of position and lexical frequency were no longer significant, and curiously, n-gram frequency which was previously non-significant now had a significant but negative effect. The combination of improved model fit and changed effects led us to consider what the random word variable was picking up. Importantly, apart from eight out of a total of 180 content word tokens that occurred twice, only function word tokens occurred more than once (range 2-74). This brought into question the rationale for entering random word effects, given that the vast majority of items occurred only once, and begged questions about the contribution of function words to the unexpected outcome.

We further explored the source of the large change in the model's predictions. Exploration of the by-Word random intercepts revealed that the definite determiner *the* had a greatly increased chance of being correctly repeated, whereas other function words (prepositions, complementizers *that/whose*, intensifier *very*, copula *be*) were some of the least accurately repeated words. In the light of this, instead of adding the by-Word random variable, we added two fixed effects (with their corresponding by-Sentence and by-Subject random variables): whether a word is the definite determiner *the* (1 = *the*, 0 = other determiner) and whether it is a function word (1 = function word but not *the*, 0 = *the* or not a function word) and fitted the model (see Table 4). The new model had predictions qualitatively identical to the model with the by-Word random variable. In line with our insights from inspecting the random intercepts, determiner *the* was much better repeated, whereas other function words were repeated less accurately than content words.

Table 4. *Estimates of fixed and random effects in the GLME model including fixed effects of determiner and function word*

Effect	Estimate	95% CI	*P*	by-Sentence SD	by-Subject SD
Intercept	2.81	2.39 – 3.23	< .0001	0.46	1.26
Determiner	2.09	1.39 – 2.79	< .0001	0.45	1.33
Function Word	-0.42	-0.73 – -0.11	< .01	0.55	0.34
Complexity	0.12	-0.32 – 0.55	= .59	-	0.66
Plausibility	0.25	0.16 – 0.33	< .0001	-	0.13
Log N-gram Frequency PC1	-0.13	-0.26 – -0.00	< .05	0.24	0.15
Log Word Position	-0.15	-0.34 – 0.04	= .13	0.4	0
Log Lexical Frequency	0.07	-0.06 – 0.19	= .29	0.23	0.14
Age	0.56	-0.18 – 1.31	= .14	0.3	-
Complexity:Age	1.06	0.51 – 1.61	< .001	-	-
Plausibility:Age	0.03	-0.08 – 0.13	= .64	-	-
Log N-gram Frequency PC1:Age	-0.11	-0.28 – 0.06	= .21	0.25	-

Discussion

This study explored the contribution of linguistic and surface distributional frequency factors to children's sentence repetition (SR) performance. While effects of syntactic complexity, semantic plausibility and n-gram frequency have been found in previous studies of sentence recall, no previous

study has sought to compare or tease these apart. Our findings revealed that, with the effects of lexical frequency and primacy taken into account, plausibility contributed independently to SR regardless of age; complexity had reverse effects according to age group (4- to 7-year-olds vs. 11- to 12-year-olds), yielding lower accuracy in the younger group and greater accuracy in the older group; and we found no evidence for an effect of n-gram frequency for either age group. However, post hoc exploratory analysis of word-level variables revealed that function word type was a significant factor and furthermore impacted on three other factors: with function word type included, effects of word position and lexical frequency ceased to be significant, while n-gram frequency now had a significant but negative effect. Age did not contribute significantly to overall accuracy. In the following sections we discuss each of these outcomes in relation to findings reported in previous studies, and explore their implications, individually and combined, for the levels of knowledge children use in the particular task of SR and for theories of language processing more generally.

Syntactic complexity effects. Our finding that young children had disproportionate difficulty with complex sentences and that accuracy in this condition improved with age is in line with previous research viewing SR as an index of morphosyntactic development (e.g. Klem et al., 2015; Polišenská et al., 2015) It is also in line with our hypothesis that abstract distributional knowledge is less available or robust in younger children but older children's repetition of complex sentences would benefit from such knowledge. This adds to evidence that complex structures are still emergent in the early primary years but well established by the early secondary years (for a review, see Ambridge et al., 2018). While this is consistent with the theory that knowledge of complex sentence structure arises through abstraction across multi-word chunks, it is contrary to purely exemplar based accounts that attribute learning to frequency of multi-word chunks.

More surprising is our finding that older children repeated complex sentences with greater accuracy than simple sentences. We speculate that knowledge of complex morphosyntactic structures in older children might serve as scaffolding that boosts their retention of parts of sentences, particularly where support from semantic knowledge is limited as was the case for our more implausible sentences. For example, *The boat that the bird wore folded a penny* contains a frequent

morphosyntactic frame 'the N that the N' which may facilitate recall of the content relative to, for example, a simple dative sentence such as *The spider from home told the shoe a candle*. Having access to this scaffolding information might free up processing/storage resources to focus on content words that fill slots in the frame. On this interpretation, familiarity with the morphosyntax of complex sentences could account for the unexpected advantage, relative to simple sentences, observed in the older group. It should be noted that frequency in this account relates to specific morphosyntactic structures/frames and the sequence of function words that define the frame (e.g. *the N that* $_{NP}$*[the…]*), rather than to frequency of n-grams. This implicates a similar level of abstraction to that invoked by Kidd et al. (2007) in their discussion of frequency effects in children's repetition of subject and object relative clauses (see Introduction), with a specific advantage for relative clauses containing *inanimate NP – that – pro*. This would be evidence for morphosyntactic knowledge, rather than exemplar-based knowledge, playing an important role in sentence repetition, and is in line with evidence that children's capacity to repeat sequences of words is dramatically increased when the sequence of words is consistent with morphosyntactic frames in the language (Polišenská et al., 2015). However, further investigation is needed to support this interpretation.

Since the effects of complexity reversed across age and there was furthermore almost no shared variance between complexity and n-gram frequency (see Table 1), we might conclude that these effects are not reducible to the n-gram frequency of simple vs. complex sentences, supporting the view that SR is sensitive to syntactic knowledge. However, this interpretation overlooks the possibility that the number of multi-word sequences (i.e. n-grams) specific to complex sentences as well as the variety of complex sentences to which children are exposed may increase with age; in this case, changes in language input and experience, rather than the consolidation of abstract structure, might account for the reverse effects of complexity in our younger and older groups, supporting an n-gram based account. Since the corpus from which we derived our n-gram metrics was not differentiated for age, we cannot rule out the possible contribution of changes in n-gram input frequency to the complexity effects we observed in our two age groups. We discuss this issue in more detail below.

Age. Our finding that age effects were confined to the complex condition was contrary to our prediction and to wide-ranging evidence of age effects on SR. Overall levels of accuracy were high (as indicated by the percentage of words repeated accurately), and we infer that our plausible sentence targets with simple syntactic structure were well within the capacity of the younger as well as the older group.

Plausibility effects. The plausibility effect we observed was as predicted and matched findings of previous studies of SR in children and adults (e.g. Miller & Isard, 1963, Polišenská et al., 2014; Wallan et al., 2011; Valian et al., 2006). In contrast to our findings for complexity, the interaction of plausibility with age was not significant, so this effect appeared to operate across childhood (4 - 12 years in this study) and this too was in line with our hypothesis. It is likely that propositions in our plausible sentences were equally familiar to younger and older children, so there was no variation across age in the conceptual-semantic knowledge supporting plausible sentences. While older children have greater experience and understanding of real world situations which contributes to their growing knowledge of semantic-morphosyntactic configurations, this knowledge did not appear to benefit sentences that contravene conceptual-semantic knowledge.

Since the violations in our implausible sentences varied from semantic incompatibility (e.g. …*comb*… *was happy*) to possible but unlikely scenarios (e.g. ...*page*… *was hot*), plausibility effects could arise from semantic and/or real world knowledge. On a semantic interpretation, plausibility effects would arise from frequency of pairings between lexical-distributional input and concepts and thematic roles. For example, children may know what noun or type of noun enters what thematic role with what predicate without necessarily accessing the message encoded by the construction. There is some evidence to support this line of thinking. For example, young children more quickly comprehend phrases with semantically related nouns and verbs compared to semantically unconstrained verbs (Fernald, 2004). Similarly, Yuan, Fisher, Kandhadai and Fernald (2011) showed that 2-year-old children were able to learn the semantic category of direct objects of novel verbs from listening experience, by having heard a novel verb in sentences with nouns of a particular category. Interestingly, having heard novel nouns used near nouns of a particular category did not produce the same learning pattern. However, it is also possible that plausibility effects are the product of real

world knowledge, arising because children find it easier to access and recall the situations encoded by plausible than implausible sentences.

Whatever the specific source of the plausibility effect is – knowledge of the world and/or semantic relations coded by verbs/predicates – our findings show the effect of plausibility cannot be reduced to knowledge of frequently co-occurring lexical items. Similarly, Jolsvai, McCauley and Christiansen (2013) found in a phrasal decision task that adult participants' reaction times were affected by lexical co-occurrence frequency but the effect of meaningfulness was even larger. Future research could construct sentences that manipulate semantic and real world knowledge (e.g., targeting semantically anomalous vs. pragmatically unusual sentences, categorically or on a continuum), and frequency of sequential but also non-sequential word co-occurrence using algorithms computing semantic relatedness based on bag-of-words algorithms (e.g., Latent Semantic Analysis, Landauer, & Dumais, 1997; or word2vec, Mikolov, Sustkever, Chen, Corrado, & Dean, 2013). This approach could shed further light on the levels of semantic-conceptual knowledge that contribute to SR and how these relate to lexical-distributional content.

N-gram effects. Contrary to our hypothesis and previous findings, we found no evidence that higher frequency multi-word chunks increased accuracy of repetition in either age group. In contrast, lexical frequency was found to have a significant effect (but see a breakdown of this effect in post hoc exploratory analyses reported below).

It is possible that design issues contributed to this unexpected pattern of results. Unlike previous studies, our study included plausibility as a predictor, and we cannot entirely rule out a confound between n-gram frequency and plausibility. However, the correlation between PC1 and plausibility was very low (see Table 1), indicating very little risk of confounding. More importantly, whereas our study took continuous measures of n-gram frequency, previous studies manipulated n-gram frequency in sequences designed to exclude other factors, resulting in highly controlled sets of stimuli (Arnon & Snider, 2010; Arnon & Priva, 2013; Bannard & Matthews, 2008; Supasiraprapa, 2019). As a result, our measure of n-gram frequency differed from previous studies in three key respects, all of which may have contributed to the lack of a positive effect of n-gram frequency in our study. First, our study considered in total 952 n-grams which varied in morphosyntactic and lexical

composition. In contrast, previous studies used a small selection of n-gram pairs which differed only with respect to the final word, which was almost without exception a content word, for example, *don't have to worry* vs *don't have to wait* (Arnon & Snider, 2010); *a lot of noise* vs *a lot of juice* (Bannard & Matthews, 2008). Final words were matched in lexical frequency, but n-gram pairs differed in frequency of the multi-word expression. Hence, while these studies showed that n-gram frequency effects occurred when other, potentially more important influences, had been eliminated, they did not reveal the extent to which frequency effects account for sentence processing performance.

Second, as illustrated by these examples, most stimuli in previous studies formed a syntactic and/or intonational phrase, whereas our study considered all n-grams in the target sentences including the many n-grams that straddled intonational and syntactic boundaries. A study from the field of second language acquisition by Ellis, Simpson-Vlach and Maynard (2008), which included n-grams straddling syntactic constituent boundaries, found that n-gram frequency did not significantly affect native speakers' performance, while n-grams with high coherence provided a processing advantage. Furthermore, some of the stimuli in previous studies not only formed a syntactic/intonational unit, but were formulaic expressions, for example *how do you do, all over the place, on the other hand.* Such items have a meaning as a unit and may be learned and entrenched as such, so we would expect them to behave like lexical items and show sensitivity to frequency. In contrast, our targets did not include any formulaic sequences. This difference raises the question of whether effects of n-gram frequency observed in previous work were driven primarily by n-grams that formed a unit (syntactic, intonational and/or semantic).

Third, although the paired items used in previous studies varied in n-gram frequency, all were at the relatively high end (note that in Bannard and Matthews' study the targets were chosen to reflect the language young children encounter and therefore had to be at the high frequency end of the spectrum). In contrast, our targets varied widely in multi-word frequency; indeed, 25% of our 3-grams had no matches at all in the large n-gram frequency corpus. We obtained n-gram frequencies for the stimuli used in Arnon and Snider (2010) and Bannard and Matthews (2008) from the Google N-Gram corpus used in our study and compared their range of n-gram frequencies with our study. We focused on the 3-grams as the most representative n-gram length (see Table 1). The average log frequency of

3-grams in the other two studies is 13.44 (*SD* = 1.95), compared with 6.93 (*SD* = 2.83) in our items. The marked difference in the distribution of 3-gram frequency is illustrated in Figure 2. It is therefore possible that n-gram frequency affects language production and comprehension only at the very high end of the n-gram frequency spectrum sampled in previous studies, and that we did not observe effects of n-gram frequency because our stimuli contained very few n-grams in this high-frequency range. As Jacobs, Dell, Benjamin and Bannard (2016) point out, novel combinations are typical of real-life language, and multi-word sequences necessarily have a lower frequency range than words. Hence the frequency of high-frequency phrases is at best similar to that of low frequency lexical items.

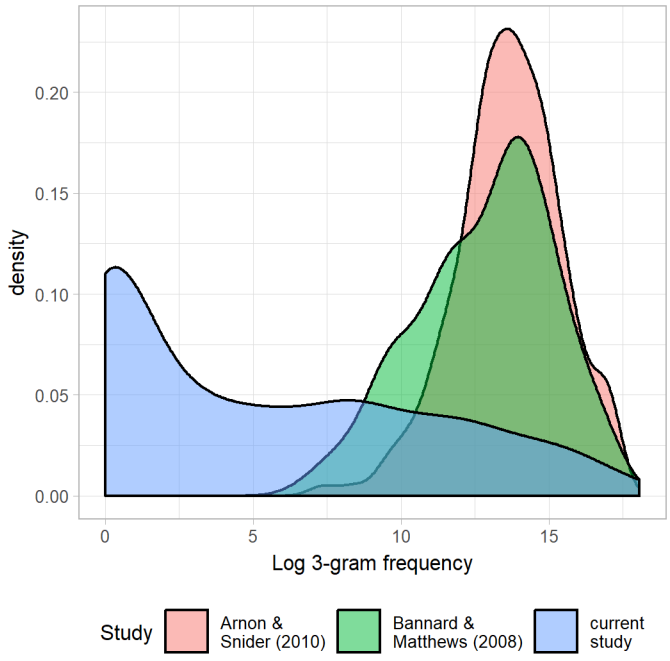

Figure 2. Comparison of n-gram frequencies of our stimuli and previous studies.

A final issue in considering frequency effects is the corpus from which n-gram frequencies are derived. It might be argued that Google Books N-gram Corpus was not an appropriate source for evaluating n-gram frequency in our study on the grounds that sentence repetition involves oral rather than written language, and that language in books may not be representative of language experienced in the spoken or written modality by children in the age ranges we sampled. This issue is a very serious challenge facing studies of frequency effects, namely, what is a relevant and representative

corpus? For children in the early stages of language acquisition, parent/carer input represents the bulk of the child's language experience, and for children of this age the CHILDES database (MacWhinney, 2000) provides a wealth of transcribed child-directed input. However by age 3 years, if not earlier, children encounter a range of interlocutors in a range of settings (e.g. nursery staff, older children), more varied modalities (not only face-to-face conversation, but shared storybook reading, and on-screen language). The ideal corpus for our purposes would be an extensive sample of input representative of the language and contexts in the current or cumulative experience of children of our participants' ages. Because to-date there exists no such corpus of language experienced by school-aged children, previous studies have also used corpora based on written language (e.g., British National Corpus: Ambridge, 2013; Ambridge, Barak, Wonnacott, Bannard & Sala, 2018; ICE-GB: Ambridge, Pine & Rowland, 2012) while studies with adults have also used Google Books (e.g., Onnis & Thiessen, 2013). An equally, if not much more important consideration, is the size of the corpus. The frequency of word n-grams (especially longer than bigrams) in the full range of frequency spectrum can only be reliably estimated using a large corpus. Here, the size of the Google Books N-gram Corpus and diversity of genres represented in books (including written representations of spoken language) made it the strongest available corpus for deriving a generalised measure of frequency of a wide range of n-grams at multiple levels of granularity.

Further insights from a post-hoc exploratory analysis. A post-hoc exploratory analysis of by-word random intercepts changed the effects of lexical and n-gram frequency. The most surprising finding was that the effect of n-gram frequency in the updated model was weakly negative, i.e. the model predicted that the more frequent the word sequence, the less accurate children's recall of the constituent words. The fact that the pattern of results changed qualitatively when the analysis controlled for new word-level variables has important methodological implications for future studies that focus on sentence processing effects at the word level. Apart from lexical frequency, other word-level predictors such as word category and the syntactic type of phrase in which the word occurs should be included. Such factors may explain a considerable amount of variance which, if unaccounted for, introduces noise into the measurement and may be confounded with the effects of interest.

In the case of our stimuli, the post hoc analysis suggested that such factors influenced the contribution of lexical as well as n-gram frequency. Our post hoc finding of a significant advantage for repetition of the definite article *the* is unsurprising. Given the typical SVO structure of English sentences, the typical [determiner noun] structure of noun phrases, and the disproportionately high frequency of determiner *the*, children are likely to include subject and object NPs containing *the* when they repeat a sentence containing these whether or not they recall the constituent nouns. Our stimuli may have strengthened the advantage for *the* because all began with a subject NP containing *the*; furthermore, *the* served as the determiner in almost all other NPs in the target sentences. So with knowledge of the role and structure of NPs in English, recall of *the* could be seen as coming "for free". Conversely, the less accurate repetition of all other function words in our stimuli could be due to the fact that almost all occurred in *optional* phrases (prepositional phrases, adjective phrases, or relative clause modifiers) which might be more susceptible to omission even though they comprise relatively high-frequency n-grams. Since all function words are very high frequency, both the positive effects of *the* and the negative effects of other function words were confounded with lexical frequency in the original analysis. Given our observations about the unique characteristics of definite determiner *the*, it seems plausible that it was the effect of *the* that carried the positive effect of lexical frequency in that analysis. Given our observations about the position and n-gram frequency of phrases containing all function words apart from *the*, it is possible that the negative effect of these function words carried the effect of position in our original analysis and the negative effect of n-gram frequency in our post hoc analysis. Because this is already a post-hoc analysis, we leave the testing of these hypothetical accounts to further studies, and simply conclude that the post-hoc analyses at the word level, invited by an anonymous reviewer, do not support claims that n-gram frequency has positive effects across a broad spectrum of n-gram frequency.

Implications for the levels of knowledge contributing to sentence repetition. To our knowledge, our study is the first to evaluate effects of frequency at multiple n-gram levels varying continuously rather than categorically[2] across a wide frequency range, relative to the contribution of semantic-

[2] Arnon and Snider (2010) designed their stimuli as binary pairs with respect to frequency (high vs low frequency n-gram stimuli), but also showed in a reanalysis of the same stimuli that when n-gram

conceptual and morphosyntactic knowledge. The outcomes of our analyses are in line with the theoretical view that compositional knowledge plays the major role in sentence repetition. Overall, the effects of complexity and plausibility, while controlling for n-gram frequency, suggest that repetition of syntactically and lexically varied targets reflects much more than sheer frequency of exposure to lexical combinations. Since plausible and implausible conditions were matched for morphosyntactic structure, and the plausibility effect was relatively independent of n-gram frequency, we can conclude that conceptual-semantic knowledge contributed to sentence repetition. Sentence complexity effects were also independent of n-gram frequency indicating the contribution of syntactic knowledge. However, the effects of syntactic complexity reversed with age. We attributed this reversal to the developmental trajectory in processing and consolidating the more complex morphosyntactic-semantic relations that complex sentences encode, suggesting that the morphosyntactic structure of complex sentences, once consolidated, provides a scaffolding that supports repetition in older children.

The effects we have observed are in line with recent computational and theoretical accounts in which sentence processing is the product of linguistic input, situational contexts and connections between these (e.g. Chang, Dell, & Bock, 2006; Chiat, 2001; Ambridge et al., 2015; Ambridge, 2019). While some computational work has successfully used surface-level distributional information to simulate early language acquisition (e.g. Freudenthal, Pine, Jones, & Gobet, 2015; McCauley & Christiansen, 2017; Monaghan & Christiansen, 2010), the current study suggests that surface distributional statistics are by no means the whole story. Our results indicate that children's SR also depends on familiarity of sentence meaning, which could arise from semantic knowledge based on semantic-morphosyntactic pairings (e.g. knowledge that the verb *eat* requires a subject that is animate and a direct object that is edible) and/or real world experience (e.g. knowledge that people eat apples but not clouds); it is also affected by syntactic complexity, which could arise from frequency of abstract morphosyntactic constructions and/or the complex computation of syntactic-semantic relations between their constituent words. The lack of positive n-gram frequency effects in our study,

---

frequency was treated as a continuous variable, it was a better predictor of reaction times than a categorical one.

in contrast to previous findings, suggests that such effects may be confined to combinations of words in the high-frequency range that become entrenched, and most extremely, in formulaic word combinations that have a non-literal meaning and behave like a single lexical item.

Conclusion. We set out to address competing theories regarding the role of abstract linguistic knowledge and surface-level knowledge in language processing. This is the first study to examine simultaneously the effects of frequency of multi-word sequences (n-grams) and linguistic factors of syntactic complexity and semantic plausibility on sentence-level processing. We investigated the contribution of these factors to sentence repetition performance of children in the early primary school and early secondary school years. Our study found that children's sentence repetition was affected by semantic information, and by syntactic complexity depending on age, but not by frequency of n-grams. Previous studies have found n-gram frequency effects in highly constrained comparisons of n-gram pairs that were limited with respect to other potentially important factors. The n-grams considered in our study extended across a wide range of n-gram frequency, varied length, and varied structural and semantic content and context, making it an ecologically more valid evaluation of n-gram frequency. Our results suggest that the effects of n-gram frequency observed in previous studies diminish and may be overridden once morphosyntactic and semantic factors come into play. It appears that children's performance cannot be explained simply by surface level lexical co-occurrence (n-gram frequency) but involves higher-level linguistic processing (syntactic and semantic knowledge). Our study does not rule out that in certain situations lexical co-occurrence facilitates sentence processing, but it does suggest that the facilitation is confined to extremely high co-occurrence frequencies and/or formulaic expressions. In this case, the use of memorized word sequences cannot be the central mechanism underlying language acquisition and processing.

Our findings offer new support for theories of children's language acquisition and language processing which move beyond the argument for purely statistically-based learning as opposed to rule-based learning, instead highlighting the dynamic, online interaction of multi-level information during acquisition, processing and production (e.g. Ambridge et al., 2015; Ellis, O'Donnell, & Römer, 2014), and raising questions about the relative contribution of different levels of information to

different language behaviours/tasks and at different ages (Bannard & Matthews, 2008; Jacobs et al., 2016; Frank & Christiansen, 2018).

References

Abbot-Smith, K., & Behrens, H. (2006). How known constructions influence the acquisition of other constructions: The German passive and future constructions. *Cognitive Science*, *30*(6), 995–1026. https://doi.org/10.1207/s15516709cog0000_61

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition. *First Language.* doi:10.1177/0142723719869731

Ambridge, B. (2013). How do children restrict their linguistic generalizations? An (un-) grammaticality judgment study. *Cognitive Science*, *37*(3), 508–543. https://doi.org/10.1111/cogs.12018

Ambridge, B., Barak, L., Wonnacott, E., Bannard, C., & Sala, G. (2018). Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis. *Collabra: Psychology*, *4*. http://doi.org/10.1525/collabra.133

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

Ambridge, B., Pine, J. M., & Rowland, C. F. (2012). Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition*, *123*(2), 260–279. https://doi.org/10.1016/j.cognition.2012.01.002

Arnon, I., & Priva, U. C. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*(3), 349-371. https://doi.org/10.1177/0023830913484891

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67-82. https://doi.org/10.1016/j.jml.2009.09.005

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning the effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241-248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

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

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar. *Explaining Language Universals*, 73–101.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711–733.

Bybee, J. (2009). *Language, Usage and Cognition*. Cambridge University Press, Cambridge.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Chiat, S. (2001). Mapping theories of developmental language impairment: Premises, predictions and evidence. *Language and Cognitive Processes, 16*(2-3), 113-142. https://doi.org/10.1080/01690960042000012

Chomsky, N. (1965). *Aspects of the theory of syntax*. Oxford, England: M.I.T. Press.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39, e62. [target article: pp. 1-19]*. https://doi.org/10.1017/S0140525X1500031X

Culicover, P., & Jackendoff, R. S. (2005). *Simpler Syntax*. Oxford: Oxford University Press.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Phil. Trans. R. Soc. B, 369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

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

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics 25*(1), 55-98. https://doi.org/10.1515/cog-2013-0031.

Ellis, N. C., Simpson-Vlach, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, *42*(3), 375-396.

Fernald, A. (2004). The search for the object begins at the verb. Presented at *The 29th Annual Boston University Conference on Language Development*, Boston, November 4-7.

Frank, S. L., & Christiansen, M. H. (2018). Hierarchical and sequential processing of language: A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: insights from neuroscience. *Language, Cognition and Neuroscience*, *33*(9), 1213–1218.

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition*, *143*, 61–76. https://doi.org/10.1016/j.cognition.2015.05.027

Frizelle, P., & Fletcher, P. (2014). Relative clause constructions in children with specific language impairment: Relative clause constructions in children with SLI. *International Journal of Language & Communication Disorders*, *49*(2), 255–264. https://doi.org/10.1111/1460-6984.12070

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224.

Hare, M., McRae, K., & Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, *48*(2), 281–303.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multi-word sequences. *Journal of Memory and Language*, *87*, 38-58.

Jefferies, E., Ralph, M. A. L., & Baddeley, A. D. (2004). Automatic and controlled processing in sentence recall: The role of long-term and working memory. *Journal of Memory and Language*, *51*(4), 623–643. https://doi.org/10.1016/j.jml.2004.07.005

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2013). Meaning overrides frequency in idiomatic and compositional multi-word chunks. In N. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 692–697). Austin, TX: Cognitive Science Society.

Kidd, E., & Bavin, E. L. (2002). English-speaking children's comprehension of relative clauses: Evidence for general-cognitive and language-specific constraints on development. *Journal of Psycholinguistic Research*, *31*(6), 599-617.

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

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science*, *18*(1), 146-154. https://doi.org/10.1111/desc.12202

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211. https://doi.org/10.1037/0033-295X.104.2.211

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford: Stanford University Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database* (Vol. 2). Lawrence Erlbaum.

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multi-word chunks in language learning. *Topics in Cognitive Science*, *9*(3), 637–652.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Pinker, S. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182. https://doi.org/10.1126/science.1199644

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing*

Systems (pp. 3111–3119). Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, *2*, 217–228. https://doi.org/10.1016/S0022-5371(63)80087-0

Moll, K., Hulme, C., Nag, S., & Snowling, M. J. (2015). Sentence repetition as a marker of language skills in children with dyslexia. *Applied Psycholinguistics*, *36*(2), 203-221. https://doi.org/10.1017/S0142716413000209

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(3), 545–564. https://doi.org/10.1017/S0305000909990511

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268–284. https://doi.org/10.1016/j.cognition.2012.10.008

Polišenská, K., Chiat, S., Comer, A., & McKenzie, K. (2014). Semantic effects in sentence recall: The contribution of immediate vs delayed recall in language assessment. *Journal of Communication Disorders, 52*, 65-77. https://doi.org/10.1016/j.jcomdis.2014.08.002

Polišenská, K., Chiat, S., & Roy, P. (2015). Sentence repetition: What does the task measure?. *International Journal of Language & Communication Disorders, 50*(1), 106-118. https://doi.org/10.1111/1460-6984.12126

Riches, N. G. (2012). Sentence repetition in children with specific language impairment: An investigation of underlying mechanisms. *International Journal of Language & Communication Disorders*, *47*(5), 499–510. https://doi.org/10.1111/j.1460-6984.2012.00158.x

Riches, N. G. (2017). Complex sentence profiles in children with Specific Language Impairment: Are they really atypical?. *Journal of Child Language*, *44*, 269-296. https://doi.org/10.1017/S0305000915000847

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language, 25*(1), 169-201. https://doi.org/10.1017/S0305000997003395

Supasiraprapa, S. (2019). Frequency effects on first and second language compositional phrase comprehension and production. *Applied Psycholinguistics*, *40*(4), 987-1017.

Szewczyk, J. M., Marecka, M., Chiat, S., & Wodniecka, Z. (2018). Nonword repetition depends on the frequency of sublexical representations at different grain sizes: Evidence from a multi-factorial analysis. *Cognition*, *179*, 23-36.

Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development*. Cambridge University Press.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*(2), 569-613. https://doi.org/10.1111/j.1467-9922.2010.00622.x

Valian, V., Prasada, S., & Scarpa, J. (2006). Direct object predictability: Effects on young children's imitation of sentences. *Journal of Child Language, 33*(2), 247-269. https://doi.org/10.1017/S0305000906007392

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190. https://doi.org/10.1080/17470218.2013.850521

Wallan, A., Chiat, S., & Roy, P. (2011). *Evaluation of an Arabic sentence-repetition test for preschool children*. Poster presented at the ASHA Convention 2011, San Diego.

Yuan, S., Fisher, C., Kandhadai, P., & Fernald, A. (2011). You can stipe the pig and nerk the fork: Learning to use verbs to predict nouns. In N. Danis, K. Mesh, & H. Sung (Eds.), *Proceedings of the 35th Annual Boston University Conference on Language Development* (pp. 665-677). Somerville, Massachusetts: Cascadilla Press.

Acknowledgments

We thank Kathryn Henry, Laura Mills, Paul Wallis and Emma Whittingham for help with data collection, and Charlotte Bown for help with transcription and scoring.

Appendix. List of stimuli

Complexity	Plausibility	Construction	Target sentence
Simple	P	S V O A	The woman found a small frog in the kitchen
	I	S V O A	The butter lost a green nurse in the window
	P	S Vdative DO IO	The clown gave the white jacket to the dancer
	I	S Vdative DO IO	The cloud showed the bad bubble to the ladder
	P	S V O A	The girl wanted a lovely toy for her friend
	I	S V O A	The cat painted a tasty cup for her swing
	P	S V O A	The bird picked a red apple from the garden
	I	S V O A	The bear pushed a cheap towel from the curtain
	P	S with modifier V O	The child in the corner wrote a long story
	I	S with modifier V O	The pen in the carrot read a deep picture
	P	S with modifier V C	The bike with a huge wheel was very heavy
	I	S with modifier V C	The comb with a new nail was very happy
	P	Dative IO DO	The cook gave the children a biscuit for tea
	I	Dative IO DO	The book made the people a whisker for fun
	P	Dative IO DO	The teacher from school showed the boy a hamster
	I	Dative IO DO	The spider from home told the shoe a candle
	P	Possessive	The mother left the key by the neighbour's door
	I	Possessive	The person kept the hair by the lady 's jam
	P	Possessive	The dog hid the teddy 's blanket in the box
	I	Possessive	The neck put the tiger 's carpet in the fish
	P	Passive	The smelly cheese was eaten by a big mouse
	I	Passive	The sleepy bus was broken by a sad house
	P	Passive	The can was opened by a very sharp knife
	I	Passive	The nose was melted by a very dry egg
Complex	P	SO relative	The train that the boy missed was very busy
	I	SO relative	The grass that the milk kicked was very silly
	P	SO relative	The man that the girl saw posted a letter
	I	SO relative	The boat that the bird wore folded a penny
	P	OO relative	The queen hated the dress that the lady bought
	I	OO relative	The tree needed the floor that the rabbit caught

P	OO relative	The monkey took the nuts that the baby threw
I	OO relative	The table hit the socks that the tummy drew
P	S Genitive relative	The driver whose coat was dirty sold the car
I	S Genitive relative	The basket whose shirt was empty liked the sky
P	S Genitive relative	The woman whose bag was full dropped the ticket
I	S Genitive relative	The teacher whose page was hot hugged the packet

Note: P = plausible sentence, I = implausible sentence, S = Subject, V = Verb, DO = direct object, IO = indirect object, O = object, A = Adverbial, C = complement