Comparing children and large language models in word sense disambiguation: Insights and challenges

Francesco Cabiddu
University College London, UK

Mitja Nikolaus
CerCo, CNRS, France

Abdellah Fourtassi
Aix-Marseille Université, France

Abstract: Understanding how children process ambiguous words is a challenge because sense disambiguation is a complex task that depends on both bottom-up and top-down cues. Here, we seek insight into this phenomenon by investigating how such a competence might arise in large distributional learners (Transformers) that purport to acquire sense representations from language input in a largely unsupervised fashion. We investigated how sense disambiguation might be achieved using model representations derived from naturalistic child-directed speech. To this end, we tested a large pool of Transformer models, varying in their pretraining input size/nature as well as the size of their parameter space. Tested across three behavioral experiments from the developmental literature, we found that these models capture some essential properties of child word sense disambiguation, although most still struggle in the more challenging tasks with contrastive cues. We discuss implications for both theories of word learning and for using Transformers to capture child language processing.

Keywords: Child Word Sense Disambiguation; Transformers; Usage-Based Learning.

Corresponding author(s): Francesco Cabiddu, UCL Psychology and Language Sciences, University College London, 26 Bedford Way, London, WC1H OAP, UK. Email: francesco.cabiddu@ucl.ac.uk.

ORCID ID(s): https://orcid.org/0000-0001-9692-4897

Citation: Cabiddu, F., Nikolaus, M., & Fourtassi, A. (2025). Comparing children and large language models in word sense disambiguation: Insights and challenges. *Language Development Research, 5*(1), 35—115. http://doi.org/10.34842/k0rg-2n19

Introduction

Large language models are deep artificial neural networks pretrained on large unlabeled datasets via self-supervised learning. These models have had a great impact in the field of Natural Language Processing (hereafter NLP) for their performance in language understanding and generation tasks (e.g., Bommasani et al., 2022). Here, we examined the plausibility of these models as distributional learners posited by usage-based approaches of language acquisition (e.g., Ambridge, 2020; Bybee, 2010). We focused on child word sense disambiguation (Cabiddu et al., 2022b; Rabagliati et al., 2013). That is, how children use sense-specific representations (e.g., band = music band, elastic band). Specifically, we examined whether the distributional learning mechanisms that allow these models to acquire linguistic knowledge at the sentence and word level could give rise to word sense disambiguation skills that children exhibit in behavioral tasks.

We tested models based on the Transformer architecture (Vaswani et al., 2017) that perform sense disambiguation using sentence context to form contextualized representations. Transformers are sensitive to both bottom-up direct word-associations (a word co-occurring frequently with another across different sentences) and top-down syntactic and semantic sentence structures (e.g., Jawahar et al., 2019; Tenney et al., 2019) on which sense disambiguation depends. Here, we refer to these high-level sentence structures as top-down cues that a usage-based learner might acquire through language experience (Alishahi & Stevenson, 2013; Bybee, 2010). These refer to any abstract knowledge that might enable an individual to generalise a certain sentence structure to novel language instances (e.g., a child knowing that “pushing a flowerpot” is more plausible than “pushing a road” even without having heard either expression before; Andreu et al., 2013). Transformers’ inherent sensitivity to top-down cues allow us to apply these models to raw naturalistic language, without having to enrich the input with external, explicit resources to provide sensitivity to such structures. For example, Alishahi and Stevenson (2013) showed how a computational learner could apply familiar verbs to novel object arguments. The model they developed was provided with various pieces of knowledge, such as the positions of syntactic arguments within sentences and the semantic characteristics of each argument. From this, it was able to generalize the prototypical semantic properties that an argument of a verb should possess (i.e., verb-event structures; for instance, “The mechanic warned the driver” is more plausible than “The mechanic warned the engine”). This finding is significant because it provides in-principle evidence that a structural aspect like verb-event structures can be bootstrapped from input. However, providing the extensive knowledge presumed to be available to the learner requires several input pre-processing steps (e.g., lemmatizing the input, identifying and recoding naturalistic sentences as verb frames, tagging semantic characteristics of each argument using an external dictionary). It remains unclear whether the same results could be achieved without such pre-implemented knowledge in the model, relying only on bootstrapping verb knowledge directly from the input. Moreover, when one wants to apply a model to raw, naturalistic language, it becomes infeasible to pre-process the input for several aspects of sentence structure that the model should be sensitive to in order to perform certain tasks, such as word sense disambiguation.

Transformers have been used to form adult-like sense representations in natural language classification tasks, and the models have been tested on their ability to pick out a target sense given the sentence context (Loureiro, et al., 2021). However, such tasks may not suitably assess model developmental plausibility as they use coherent test sentences (i.e., all cues in the context unambiguously point toward one target sense). Relying on these tasks makes it difficult to differentiate whether Transformers exhibit rather adult-like or child-like performance, as both adults and children have been shown to perform well at disambiguating coherent sentences (e.g., Khanna & Boland, 2010; Rabagliati et al., 2013). Thus, the goal of the current study is to test models on contrastive tasks alongside coherent ones. Contrastive tasks put bottom-up (i.e., word associations) and top-down sentence cues in competition. They represent a more suitable test of developmental plausibility because differences exist in how children and adults behave in such tasks. In fact, in sense disambiguation children rely more on bottom-up aspects of sentence context (e.g., word associations) than adults, with less reliance on top-down cues likely due to differences in language experience or slow cognitive maturation (Khanna & Boland, 2010; Rabagliati et al., 2013).

Previous studies in NLP have computed models' representations based largely on adult language (Loureiro et al., 2021, 2022). These representations are created by using a technique that computes an average representation of a word sense given a collection of sentences (e.g., a prototypical representation of a music band). Here we apply this technique to evaluate how properties of child sense processing could be captured using sense representations formed from naturalistic *child-directed* utterances. This choice is motivated by the fact that differences in how senses are assigned to words in children and adults is likely influenced by differences in word use in child and adult environments (Meylan et al., 2021). We note that this method does not involve pre-training the models on child-directed language, although we do also include a family of models pre-trained on child-directed utterances. We show that computing child-directed sense prototypes has different benefits for capturing child performance, but we also return to its limitations in the Discussion.

We evaluated Transformers using behavioral studies that tested 4-year-old children's abilities to use bottom-up (word associations) and top-down (sentence global plausibility, verb-event structure) cues to sense disambiguation (Cabiddu et al., 2022b; Rabagliati et al., 2013). We tested a large pool of models (*N* = 45) from 14 different families. This integrative approach (see also Schrimpf et al., 2021) would allow us to study how different properties of the models may lead to different behavioral patterns, while relying on a single model could be misleading as any conclusion might be influenced by idiosyncratic aspects of this specific model (architecture, pretraining objectives, amount/type of pretraining input, etc.). Specifically, we explored how scalability in models' size (number of parameters) and pretraining data size related to sense disambiguation performance. It has been shown that increasing the number of model parameters improves models' ability to generalise, enabling them to tackle a broad spectrum of language and reasoning tasks without necessitating extensive examples during training or specific model fine-tuning (e.g., Brown et al., 2020; Chowdhery et al., 2022). Essentially, more parameters in language models means a greater capacity to store patterns and nuances from the training data. This capacity to capture a wider array of linguistic patterns may lead to improved performance in tasks such as sense disambiguation, where understanding context and subtle differences in meaning is crucial. Based on findings about word age of acquisition norms (Laverghetta Jr & Licato, 2021), we expected models with a larger number of parameters to better fit child data, also in line with NLP studies showing how increasing a model's parameter count improves its ability to track both bottom-up and top-down aspects of sentence structure (Devlin et al., 2019; Hewitt & Manning, 2019; Radford et al., 2019). Similarly, better performance and generalisation abilities can be achieved by training models on larger and more diverse datasets (e.g., Raffel et al., 2023). Training models on linguistic contexts that encompass a wide range of topics, styles, and structures increases the opportunities to abstract general schemas from the linguistic examples observed. Nevertheless, there is also evidence of small (i.e., more realistic) pretraining input being enough to align models to adult neural data and reading comprehension scores (Hosseini et al., 2022), therefore we might expect a null effect of pretraining size when attempting to capture human performance.

In summary, both model size and pre-training size are dimensions that have been linked to models' generalisation abilities. This capacity is crucial for learning top-down sentence structures that can then be generalised to new linguistic instances, which is something we focus on in our study. In the following, we first introduce evidence of child sense disambiguation. Secondly, we discuss the theoretical significance of Transformers and introduce a recent framework for evaluating models in sense disambiguation.

Child Word Sense Disambiguation

Sentence context plays a significant role in sense disambiguation (e.g., Sophia [played in / twisted] a band). Children use a similar (though lower) diversity of senses in naturalistic conversations (Meylan et al., 2021), which raises a question about which sentence properties facilitate child word disambiguation (Cabiddu et al., 2022b; Hahn et al., 2015; Khanna & Boland, 2010; Rabagliati et al., 2013). Children should access cues at different linguistic levels to successfully disambiguate senses. Here, we focused on key studies that showed that 4-year-old children could use both bottom-up and top-down disambiguation cues, although to different degrees depending on the specific cue. Table 1 shows an overview of the three experiments we consider. A general goal across experiments was to test children's sensitivity to sentence context for sense disambiguation. Further, they tested if top-down cues (global plausibility, verb-event structures) played a role beyond bottom-up word associations (when the two types of cues are in direct competition). Similarly, here we investigate if Transformers could use sentence context for word sense disambiguation like children, and if they would demonstrate comparable sensitivity to top-down cues in contrastive conditions.

Table 1. ***Behavioral experiments. Target words are shown in bold. Underlined text indicates cues to the dominant sense "elastic band", while italicized text refers to cues to subordinate "music band". The Dominant selection column indicates average dominant sense selections in children, for dominant-plausible (underlined) and subordinate-plausible conditions (italicized).***

Study	Cue type	Example	Dominant selection
(Rabagliati et al., 2013) Exp 1, Coherent cues	Prior Context	Dora [looked in her drawer / *heard some music*]. The **band** was cool	79% / *33%*
	Current Context	Dora was in her room. She [stretched / *listened to*] the **band**, which was cool.	81% / *38%*
(Rabagliati et al., 2013) Exp 2, Contrastive cues	Global Plausibility	Elmo and his class were singing songs. The teacher could play music with [anything / *anyone*], even a **band**.	39% / *21%*
(Cabiddu et al., 2022b) Contrastive cues	Verb-Event Structure	Sophia listened to some music. Then she [twisted / *played in*] a **band**.	62% / *26%*
	Verb-Lexical association	Sophia listened to some music. Then she [got / *played in*] a **band**.	60% / *26%*

The behavioral studies we considered have not only been used to test children's disambiguation skills at a certain point in development but also to examine different hypotheses on whether young children can rely on the same cues for sentence parsing as adults do, or whether there are limitations in their access to certain cues that require higher levels of linguistic analysis. One account posits that children rely solely on bottom-up cues in sentence parsing (Snedeker & Yuan, 2008), while another account emphasizes cue informativity (Trueswell & Gleitman, 2007). In the context of sense disambiguation, an informativity account would suggest that children gradually refine their estimation of the general reliability of each cue (whether bottom-up or top-down) in determining word meaning as they grow. Such gradual fine-tuning could account for the differences in how children and adults perform word sense disambiguation.

The evidence available so far supports an informativity account, showing that children rely on both top-down and bottom-up cues. However, their use of top-down cues is contingent on the strength of that cue's influence in the child's early processing. For instance, children primarily rely on bottom-up word associations instead of using top-down global plausibility at the discourse level, which is the strategy predominantly used by adults (Rabagliati et al., 2013). This likely occurs because word associations are a cue that is consistently present in children's language input, and they can use this cue from very early in development. Nonetheless, this does not imply that children cannot use top-down cues. In fact, when considering a top-down cue that children also consistently use in sentence and word processing from early in development, such as verb meaning, they indeed demonstrate the ability to rely on this cue in sense disambiguation over bottom-up word associations (Cabiddu et al., 2022b).

In all studies, children heard short stories ending with a target word and saw four pictures. Two depicted the target word's alternative senses: One frequent in child-directed speech (dominant = elastic band) and one less frequent (subordinate = music band), with a 3:1 frequency ratio. The other two pictures depicted semantic distractors (e.g., sock, sport team). After the story, children chose the picture that best matched the story's final word.

In a first experiment, Rabagliati et al. (2013) tested if children could use sentence context to disambiguate dominant and subordinate senses. Disambiguation cues were presented in a previous sentence (Prior context), or in the same sentence as the target (Current context). Example stimuli are shown in Table 1. Children showed successful disambiguation across conditions, selecting more dominant senses (above 50% chance) in dominant-plausible conditions, and more subordinate senses in subordinate-plausible conditions (i.e., less than 50% dominant selections).

However, in this experiment, children could have relied solely on bottom-up associations. For example, in *Dora was in her room. She stretched the band,* one could track the association between *stretching* and *elastic band* in naturalistic conversations without processing sentence structures (i.e., using verb-event knowledge to infer that stretchable entities are usually objects). In the second experiment from Rabagliati et al. (2013) and in the experiment from Cabiddu et al. (2022b), bottom-up and top-down cues were in competition. Stories always began with a prior context containing word associates of the target subordinate sense. As shown in Table 1, prior contexts contain the words *music* or *songs* pointing toward the subordinate *music band*. Further, in experimental conditions, stories ended with top-down cues pointing toward the opposite dominant sense *elastic band* (see underlined cues in Table 1).

In Rabagliati et al. (2013) experiment 2, experimental stories shifted global semantic plausibility toward the dominant sense. Children struggled to use global plausibility and relied heavily on word associations (39% dominant selections, below chance). In other words, children struggled to use real-world knowledge, which facilitates the comprehension of causal relations, event sequences, and social norms conveyed by the overall discourse. For example, when interpreting a sentence like *Elmo and his class were singing songs. The teacher could play music with anything, even a band* the listener would need to infer that any object could emit sound and therefore, could potentially be used as a musical instrument. In contrast, children relied mostly on bottom-up word associations (i.e., tracking co-occurrences between words) to perform shallow processing of sentence context when interpreting ambiguous words (i.e., mostly interpreting *band* as a *music group* because of its association with the words *singing, songs,* and *music*).

Still, a significant difference from a control condition emerged (21% dominant selections when the story fully supported the subordinate; see italicized cue in Table 1). This result indicated residual sensitivity to top-down global plausibility in 4-year-old children.

The study by Rabagliati et al. (2013) also highlighted the limitations of capturing adults and children's reliance on top-down cues when using a distributional computational learner that is uniquely based on tracking bottom-up word associations. They employed a bag-of-words Bayesian classifier, trained on child-directed speech, to simulate children's performance in both non-contrastive and contrastive tasks. They found that while the classifier could successfully resolve non-contrastive tasks and capture variations in child performance (experiment 1), it failed in contrastive tasks (i.e., performance at floor in experiment 2, with 0% dominant senses selected across conditions), likely due to its inability to incorporate sentence-level top-down cues in its word representations. Here, we aim to examine whether a distributional learning Transformer architecture, which has shown sensitivity to top-down sentence-level structure, could instead succeed in capturing child disambiguation performance in contrastive tasks.

Cabiddu et al. (2022b) focused on verbs. Verbs are likely to represent a particularly valid cue that young children can rely on when processing sentences and words. For example, verbs' syntactic arguments guide 3- to 5-year-old children's interpretation of ambiguous sentences (e.g., Kidd & Bavin, 2005; Snedeker & Trueswell, 2004; Yacovone et al., 2021). Further, the semantic restrictions that verbs impose on their arguments (i.e., verb-event structures) guide children's unambiguous word processing (Andreu et al., 2013; Mani et al., 2016). For example, 3-year-olds know that *pushing a flowerpot* is more plausible than *pushing a road* even if they have never heard either expression (Andreu et al., 2013).

As shown in Table 1, in a Verb-Event condition, stories ended with verbs that never co-occurred with dominant senses in naturalistic conversations (i.e., children never or rarely hear *twisting a band,* which controls for verb-object associations). However, the verbs' event structure only accepted the dominant senses (i.e., one can only twist an elastic band, not a music band), making it the only available cue.

Further, the researchers examined the effect of verb-object associations (see Verb-Lexical condition in Table 1): Verbs had a neutral verb-event structure (e.g., one could get either an elastic or music band), but often co-occurred with dominant senses in naturalistic conversations (i.e., children frequently hear *getting an elastic band*). Given the role of verb-object associations in children's word processing (Mani et al., 2016), this condition tested if children would weigh more word associations coming from a verb than the rest of the (prior) context.

Children successfully resolved dominant senses using both verb-event structures and verb-object associations, beyond bottom-up word associations from prior contexts.

Overall, results from these behavioral experiments show that children can rely on different bottom-up and top-down cues for sense disambiguation. However, it remains unclear which learning *mechanisms* might underlie these competencies. Below, we use Transformers as a scientific tool to test the extent to which purely distributional learning mechanisms account for the acquisition of word sense knowledge that is dependent on sentence context.

Word Sense Disambiguation in Transformers

Testing a usage-based learner requires an architecture that forms top-down abstractions while accounting for effects of bottom-up statistical cues in language development (e.g., Ambridge et al., 2015; McCauley & Christiansen, 2019; Saffran et al., 1996). Consider the meaning of *table* in Ambridge (2019). A fixed top-down rule defining a *table* category (e.g., has legs; used for eating; made of wood, metal, or plastic; waist height) becomes falsifiable by counterexamples (e.g., an empty barrel used as a table at a bar). A solution is to embed specific contexts in the *table* representation (Ambridge, 2020; Srinivasan & Rabagliati, 2021). Bottom-up context-dependent information allows the child to estimate the similarity between a new instance *barrel table* and previously encountered *tables*. This recursive process of estimation facilitates the emergence of a context-independent, fuzzy, and probabilistic category of *table* (i.e., a prototype). In sense disambiguation, context-dependent and context-independent representations could gradually lead to multiple sense categories for a single word (Srinivasan & Rabagliati, 2021), with clusters of instances sufficiently separated in the semantic space (e.g., an object band prototype, a music band prototype).

The way sense representations are conceptualized in these proposals of word sense acquisition aligns with the ideas proposed in accounts of word sense processing (Duffy et al., 2001; Rodd, 2020). For instance, the recent semantic-settling account (Rodd, 2020) assumes that word senses are stored in a lexical-semantic space as high-dimensional representations. Distinct senses of a word form are represented as different paths embedding a set of dimensions or features that define the mapping between the word form and each sense. During sentence parsing, a settling process guides access to specific word senses by increasing the activation of specific paths in the lexical-semantic space. This activation depends on multiple cues at the word and contextual levels, helping the system settle on one sense, from bottom-up cues (e.g., meaning expectation based on words frequently co-occurring in the sentence context) to top-down cues (e.g., real-world knowledge used for pragmatic inferences). Computational evidence supporting this processing account largely comes from adult disambiguation studies (e.g., Rodd et al., 2004). However, it is still unclear whether its predictions can extend to child processing.

The above ideas of context-dependent sense representations align with Transformers' core self-attention mechanism. For each token, these models construct distinct representations that dynamically integrate sentence context. Although children have access to referential and social cues beyond sentence context, using Transformers is useful to answer the question: *How far can a distributional learner that uniquely processes naturalistic sentence context go?*

After training, Transformers encode generalized (context-independent) knowledge. Tokens from different senses organize into separate clusters within model layers, reflecting the organization of senses in dictionaries and adult representations (Loureiro et al., 2021, 2022). In Loureiro et al. (2021), Transformers were evaluated using a nearest neighbor approach (e.g., Melamud et al., 2016; Peters et al., 2018). This uses sense-annotated corpora to create model sense prototypes by averaging the representations of a collection of tokens belonging to a specific sense (see Method). Sense prototypes are then used to evaluate the model disambiguation at test. Using this method led to a Pearson's correlation of .9 between the best model and adult annotators. This method is useful because it investigates knowledge of models that are not pretrained on disambiguation, but only on predicting a word given its context (which should be more in line with what children do). Further, compared to previous studies (Haber &

Poesio, 2020), Loureiro et al. (2021) showed that models' performance better aligned with adults' when a reference sense-annotated corpus reflected the coarse-grained knowledge that adults have (e.g., collapsing senses that adults likely do not distinguish, but that are differentiated in a dictionary). This suggests that it is possible to tailor the models' sense prototypes to a specific population. In our work, reference sentences were transcribed child-directed utterances, reflecting children's naturalistic input and containing senses known to 4-year-olds based on behavioral evidence.

Method

Models

We used 13 Transformer-based language model families with varying training tasks and input encoding mechanisms. We also included a bidirectional recurrent neural network (ELMo, Peters et al., 2018), which achieved state-of-the-art results in sense disambiguation before the introduction of Transformers (e.g., Wiedemann et al., 2019). Model descriptions can be found in Appendix S1. We also share materials and code to reproduce the study results on our GitHub page (https://doi.org/10.5281/zenodo.8200803). In various configurations within families, we varied model size (number of million parameters, *M* = 287, *range* = 8 - 1,630) and pretraining size in gigabytes of text (*M* = 103, *range* = .005 - 806). In Appendix S3, we also include results from models with randomly initialized weights, showing that performance differences were not due to architectural differences in connection patterns among units.

Model Evaluation via Nearest Neighbor

Following Loureiro et al. (2021), we extracted sense prototypes using annotated sentences (see Corpora for details) in which a word occurred in a specific sense (e.g., *elastic band* in "*when we put the rubber bands around it then we'll put your name on it so we'll know which one belongs to who*"). We extracted a model's contextualized vector for each sense occurrence, summing the last four layers. For models that work at the subword level, we first averaged representations of subword tokens for the target word. Finally, we averaged the word vectors to obtain a centroid representing the *elastic band* prototype. We repeated the process for the alternative *music band.*

In Appendix S2, we also repeat the sense prototype extraction with different random samples of sentence exemplars to provide evidence that using a Nearest Neighbor approach is not heavily dependent on the specific set of exemplar sentences we used for each target sense. This decreases the concern that results from our simulations might be related to the quality of the prototypes rather than the model representations of sense usage.

To evaluate model performance, we extracted a contextualized vector for each test sentence's target word. We used cosine similarity to compare each vector with the two prototypes representing the dominant (e.g., *elastic band*) and subordinate (e.g., *music band*) senses. The most similar prototype determined the assigned sense for the test word. We then transformed this binary measure (*Dominant* = 1, *Subordinate* = 0) into a continuous measure by computing the percentage of dominant senses assigned in a specific condition (matching the child outcome measure in Table 1).

Corpora

We took sentences for computing prototypes from ChiSense-12 (Cabiddu et al., 2022a), which contains speech directed to children up to age 4 from the English section of the CHILDES database (MacWhinney, 2000). Each sentence was tagged for occurrences of 12 ambiguous words in dominant or subordinate senses (e.g., *chicken animal*, *chicken food*). The selection of dominant and subordinate senses within the corpus drew from those used in the experiments conducted by Rabagliati et al. (2013). This approach guaranteed that the chosen senses are familiar to children, as evidenced by their performance in experimental tasks. We used 9 words, excluding homophones with different spelling (e.g., son/sun) for which no ambiguity exists as the models process orthographic input. We also tagged 4 new words to cover more items from children's experiments. The target words used were all concrete nouns: band (binding or fastening object / music group); bat (animal / sports equipment); bow (knot / weapon); button (device to control electronic operations / fastener on clothing); chicken (animal / meat); glasses (eyewear / drinking vessels); letter (alphabetical symbol / mailed communication); line (geometric line / sequence of people or things arranged one behind the other); nail (body part / metal fastener); fish (animal / meat); lamb (animal / meat); turkey (animal / meat); card (playing card / greeting card).

Details about items and annotation process are in Appendix S2. The final corpus had 15,901 sentences for 13 target words, with dominant senses appearing 69% of the time on average (3:1 dominant/subordinate ratio).

Comparing Child to Model Performance

We computed an optimal outcome measure comparing child and model performance. We examined if the models exhibited a dominant sense bias reflecting the dominant/subordinate ratio in the input. For experiment 1 in Rabagliati et al. (2013) with non-contrastive cues, we fitted a linear mixed-effects model using the percentage of dominant senses selected by each model as the outcome, and model size and pretraining size as the predictors. Model family was used as random effect intercept. The model output is reported in full in Appendix S4. Only pretraining size negatively predicted dominant selection (β = -1.53, *95% CI* = [-2.30, -.75], p < .001), but not model size (β = -1.47, *95% CI* = [-3.01, .08], p = .062). As shown in Figure 1, the models better approximated the 69% dominant sense bias as pretraining size decreased.

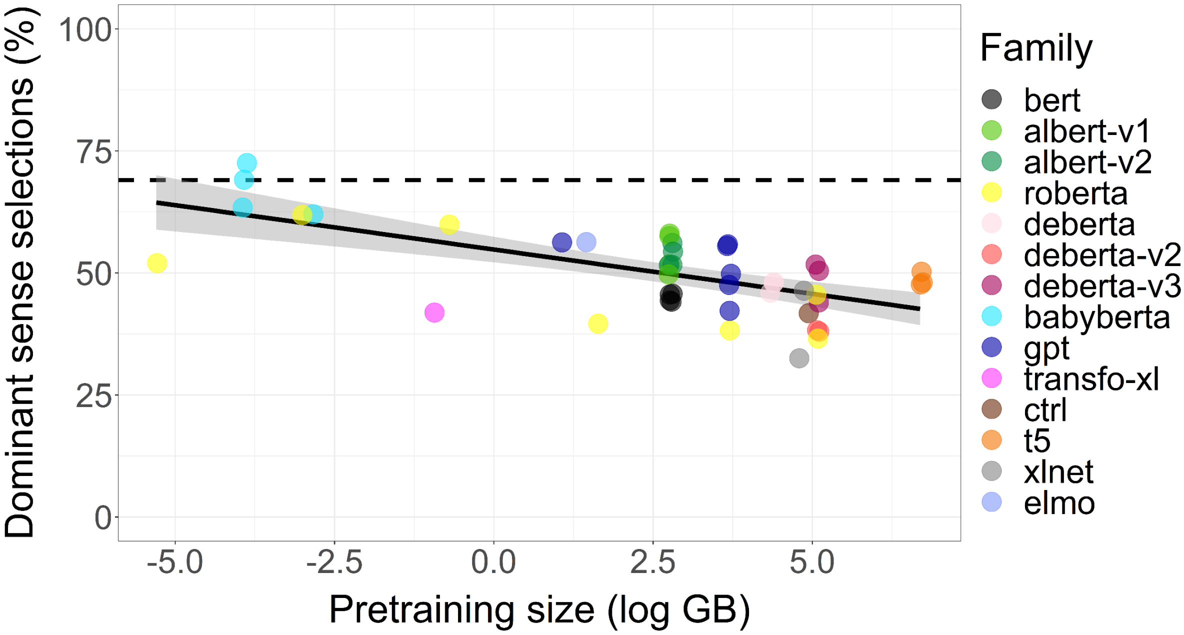

Figure 1. ***Percentage of dominant senses selected by each model in Rabagliati et al. (2013) experiment 1, by pretraining size in log GB. The dashed horizontal line indicates dominant sense prevalence in ChiSense-12.***

To further confirm that the dominant sense bias was produced by the employment of child-directed input, in Appendix S6 we also examined the models' dominant sense preference using sense prototypes computed from adult-directed speech (from utterances included in the British National Corpus; BNC Consortium, 2007). When we used adult sense prototypes, the models never approached the 69% dominant sense bias, showing equal preference for dominant and subordinate senses (50% dominant sense selections). Overall, these preliminary investigations on the effect of input speech on sense representations indicate that the use of child-directed input aligns models with children's representations of sense frequencies in naturalistic speech.

Dominant sense bias is one of the variables that can influence word disambiguation. It is an important aspect of how children disambiguate words, as well as being crucial in a model learner. However, it is not the primary focus of our examination. We aim to determine whether models are successful because they resolve the meaning of an ambiguous word using the context of the surrounding sentence, rather than from the frequency of the sense itself. For this reason, the contribution of sentence-level factors (e.g., verb information) was disentangled, for example, in Cabiddu et al. (2022b), from the contribution of word-level information (sense dominance) by statistically controlling for the latter. This was done to test whether children were indeed using verb information to resolve ambiguities. We adopted a similar approach with Transformers, by effectively separating the contribution of word-level information from that of sentence-level information. In other words, differences in dominant sense bias pose a confound: A model pretrained on a small corpus might select a similar percentage of dominant senses to children not only due to context cue sensitivity, but also because it prefers dominant senses more than a model pretrained on a large corpus. We controlled for this confound by examining the relative difference in dominant sense selections between dominant-plausible and subordinate-plausible conditions. In Appendix S5, we also include analyses that examine which models better capture children's performance when all levels (sentence-level and word-level) are considered. We return to these additional results in the Discussion.

We use relative differences in performance to control for the effect of dominant sense bias. For example, in the first experiment, children selected dominant senses (e.g., *elastic band*) in 81% of trials in the dominant-plausible condition (*She stretched the band*) and 38% in the subordinate-plausible (*She listened to the band*). For a relative difference of 81% - 38% = 43% in children, a model with 60% - 17% difference and one with 80% - 37% were considered equally similar to children. Essentially, the relative difference focused on a model's sensitivity to shifts in sentence context and compared it to children's sensitivity. The final outcome measure estimated the distance between model and children (e.g., [60% – 17%]) – [81% – 38%]), with values of 0 indicating equal sensitivity in the model and children, and values lower and higher than 0 indicating lower and higher sensitivity, respectively. Using this measure of relative distance as the outcome, we performed model comparison for each experiment between multiple nested linear mixed-effects models, which are reported in full in Appendix S4. We examined the main and interaction effects of model size and pretraining size, and employed model family as a random effect intercept in every statistical model.

Results

Rabagliati et al. (2013) - Experiment 1

Figure 2 shows models' performance by model size (2a) and pretraining size (2b). Some models reached child baseline ($y = 0$), while others performed worse ($y < 0$) or better ($y > 0$). The best linear mixed-effects model indicated higher context sensitivity as model size increased ($\beta = 5.36$, *95% CI* = [2.07, 8.64], $p = .002$) and pretraining size increased ($\beta = 3.81$, 95% CI = [2.16, 5.47], $p < .001$). A main effect of condition ($\beta = -9.98$, *95% CI* = [-16.18, -3.78], $p = .002$) showed models performing better in the current-context condition, which may not align with child performance. Although the main effect of condition was not tested in the child experiment, children's average scores might suggest similar sensitivity to prior and current context (see Table 1).

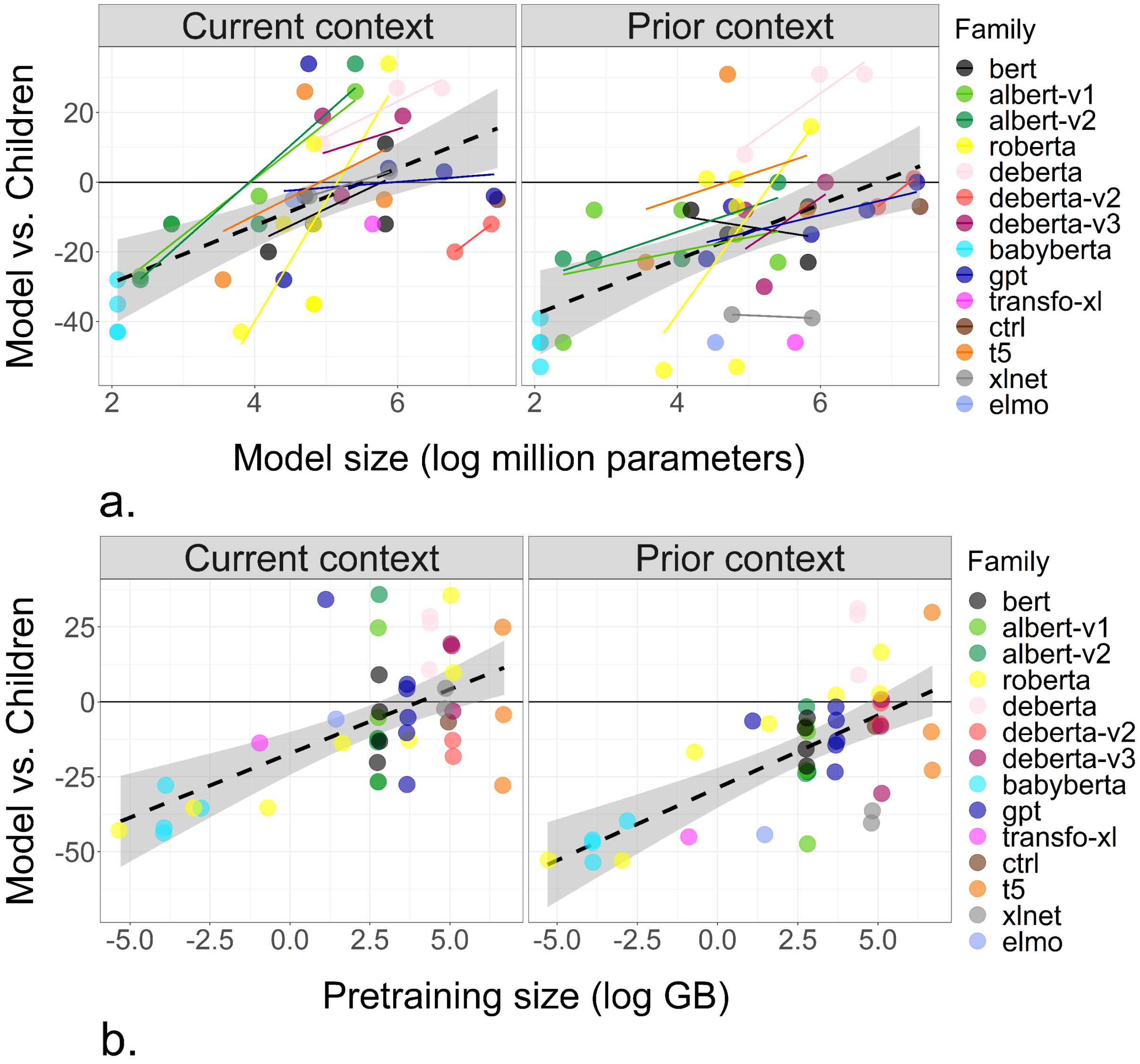

Figure 2. ***Models' relative distance from children by model size (a) and pretraining size (b), in current and prior context conditions. Model families are shown in the legend. The black horizontal line indicates child performance. The dashed regression line with 95% confidence interval shows performance across models. Colored regression lines are also shown for each model family, although only when***

examining model size as there is almost null variation in pretraining size within family. Points in panel b are jittered by 2 points in the y axis to facilitate visualization of overlapping points.

Rabagliati et al. (2013) - Experiment 2

This task used contrastive bottom-up and top-down cues, which most models seemed to struggle with: Figure 3 shows a floor effect, which led to null effects of model size (*β* = 3.37, *95% CI* = [-.35, 7.09], *p* = .075) and pretraining size (*β* = 0.12, *95% CI* = [-1.74, 1.98], *p* = .895). As confirmed in Appendix S4 (see plots showing raw dominant selection scores for each model), the floor effect led to only few models showing a difference in dominant selection between conditions. This aligns with children's residual sensitivity to top-down cues, as they displayed a difference between conditions despite low selection rates. Nevertheless, most models performed worse than children, suggesting an overall difficulty in managing contrastive cues.

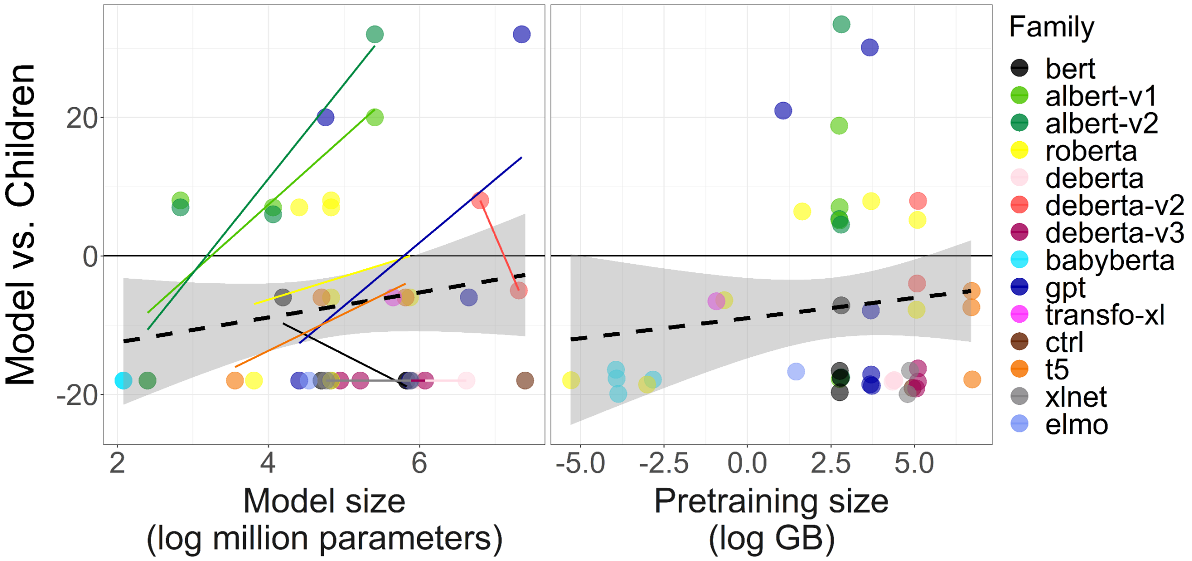

Figure 3. ***Models' relative distance from children by model size and pretraining size, in Rabagliati et al. (2013) experiment 2.***

Cabiddu et al. (2022b)

The models better handled contrastive bottom-up and top-down cues in this task, resembling the strong role of verbs in child processing. The models showed higher sensitivity to verbs with a strong event structure (Figure 4a; e.g., *She twisted a band*), with model size being positively related to models' sensitivity to verb-event cues ($\beta = 7.57$, *95% CI* = [3.48, 11.67], $p = .001$), but not pretraining size ($\beta = -.30$, *95% CI* = [-2.35, 1.74], $p = .765$). Instead, sensitivity was lower to verbs that were only lexically associated with the dominant sense (Figure 4b; e.g., *She got a band*), with no significant effects of model size ($\beta = 1.73$, *95% CI* = [-0.87, 4.34], $p = .186$) or pretraining size ($\beta = 0.16$, *95% CI* = [-1.14, 1.45], $p = .809$).

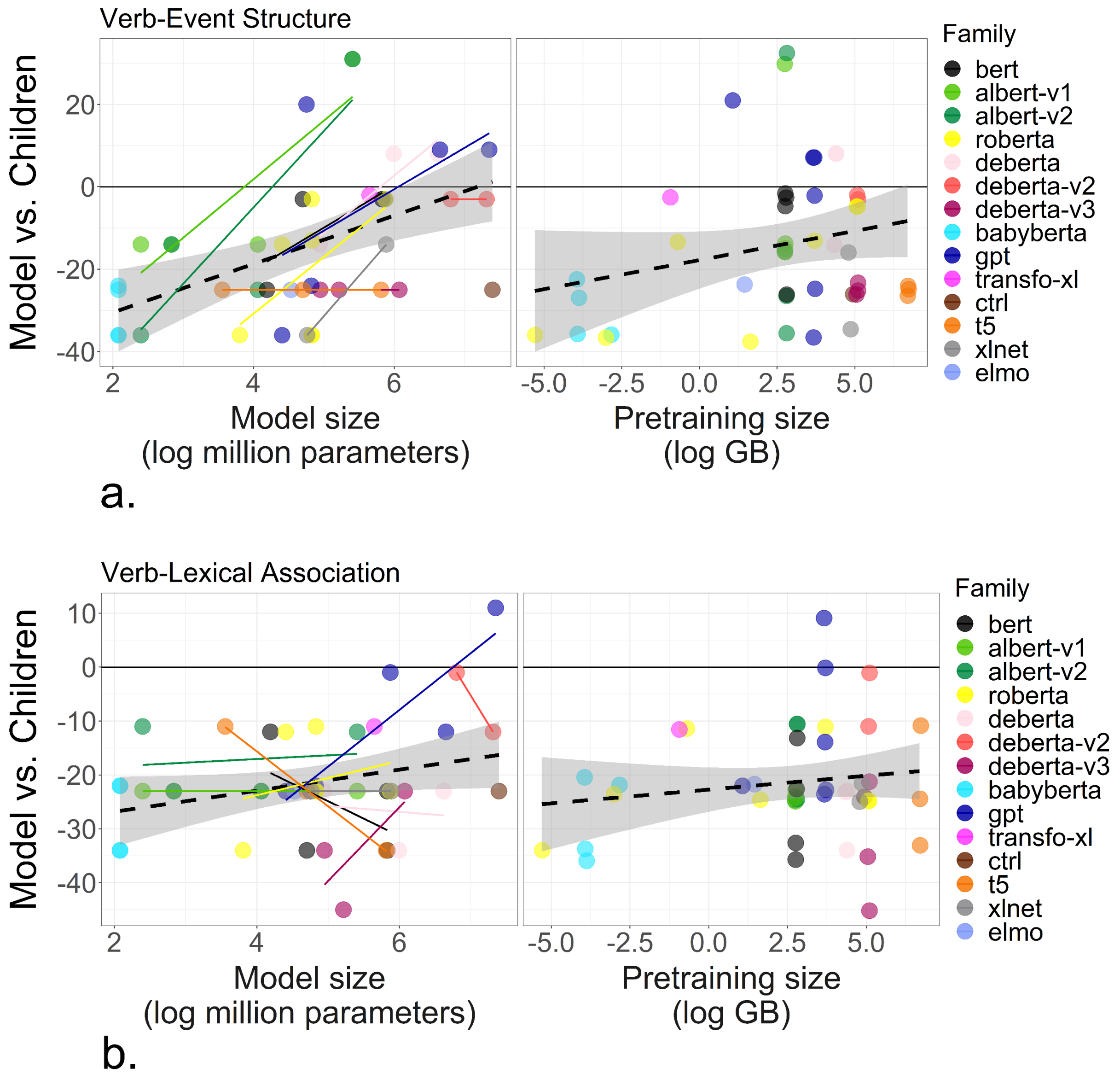

Figure 4. ***Models' relative distance from children by model and pretraining size, when examining performance at the verb-event (a) and verb-lexical conditions (b).***

Discussion

We examined the capabilities of large Transformer models in capturing child word sense disambiguation. Our results support the idea that children, like these models, might be usage-based learners who bootstrap word knowledge from the naturalistic environment (Bybee, 2010), and that child sense knowledge can, in principle, arise from probabilistic representations embedding context-dependent and context-independent information (Ambridge, 2020; Rodd, 2020; Srinivasan & Rabagliati, 2021). In line with a cue informativity account of child processing (Trueswell & Gleitman, 2007), Transformers captured the changes in word sense disambiguation performance observed across child behavioral experiments. Coherent tasks were resolved with greater ease, and performance on contrastive tasks was found to be dependent on the type of top-down cue provided (i.e., as observed in children, verbs provided a better facilitation for sense disambiguation than global plausibility).

In line with Laverghetta Jr and Licato (2021), larger models were more sensitive to both coherent (Figure 2) and contrastive cues (Figure 4a), likely because they form more precise representations based on both bottom-up and top-down aspects of sentence structure (Devlin et al., 2019; Hewitt & Manning, 2019; Radford et al., 2019).

Contrary to our prediction, models trained on larger corpora were more sensitive to coherent cues (Figure 2), while we found the predicted null effect of pretraining for contrastive cues (Figure 3 and 4). In coherent sentences, a model can rely on both word associations and top-down cues, with more pretraining likely increasing sensitivity to both. However, more pretraining might not always be as valuable for resolving *contradicting* bottom-up and top-down cues in the other conditions. Larger models might instead have an advantage in this regard.

Further, a visual inspection of models' performance at contrastive tasks (see raw plots of dominant sense selection for each model in Appendix S4) showed a stronger overall preference for subordinate senses across conditions compared to children, which might indicate models' higher sensitivity to prior context word associations (an analysis of relative differences could not highlight this, as it specifically controls for absolute differences in sense selection). In a follow-up analysis (see Appendix S5), we found evidence for this interpretation. We used an alternative outcome measure (Euclidean distance) which, compared to the relative difference, additionally looked at how close models got to $y = 0$ (Figure 2, 3, and 4) and at the exact match between models and children (i.e., difference in absolute scores): Given 81% - 38% as the children's response difference, a model performing 80% - 37% would be now closer to children than one that performs 60% - 17%. This measure might suffer from dominant sense bias (Figure 1), which we included as covariate in the statistical models to control for its effect. We replicated the positive effect of pretraining size in experiment 1 (Figure 2b), and found a *negative* effect of pretraining size in the verb-event structure condition of the third experiment (Figure 4a).

This result might indicate that smaller pretraining prevented an extreme sensitivity to word associations, allowing models to find the right balance between bottom-up and top-down cues. Interestingly, the best models in this condition received pretraining that was judged as psychologically plausible in previous studies (100 million tokens, Hosseini et al., 2022), although for an older population than ours (10-year-olds). To gain deeper insights into word association sensitivity, future work should explore how pretraining size influences the ability of large language models to track word associations and whether smaller, more realistic input can better capture children's sensitivity to these associations. Ideally, to answer this question, one would need access to the original corpora used for pretraining, which is not always possible. This would enable an understanding of precisely what types of word associations the models might have encountered during pretraining. Some recent investigations have revealed that sensitivity to word associations begins to decrease at around 1 billion tokens of input (Zhang et al., 2021). This finding might suggest the necessity to scale down to a much smaller input to avoid extreme sensitivity to bottom-up cues and to better align with child performance.

Only models with small pretraining approximated the dominant sense bias in the child input (Figure 1), and only few models (Figure 4b) showed sensitivity to verb-sense associations (e.g., *get-elastic band*), which are idiosyncrasies of the child input. One way to better align models with the child environment would be pretraining directly on child input (Hosseini et al., 2022; Warstadt & Bowman, 2022). This would also enhance the psychological plausibility of the models, which are currently pretrained on vast amounts of input, often sourced from unknown corpora and adult-directed written language. However, this task is limited by the lack of sufficiently large corpora. For example, in our study we included BabyBERTa (Huebner et al., 2021), which despite being pretrained on child input showed no sensitivity to sentence context, likely due to its small pretraining (5 million tokens). To address this gap, there is an ongoing effort within the research community to optimize model pretraining given an input limited in size, aligning more closely with human development (Warstadt et al., 2023). Model optimization also means that researchers will be able to examine and manipulate more fine-grained model dimensions than those we have considered (number of epochs, learning rate, batch size, etc.), allowing researchers to work with architectures that are likely to better approximate child learning and processing. Manipulating aspects of models' architecture will also give the opportunity to causally test their impact on the model's ability to capture child performance. For example, ablation analyses (e.g., removing parts of the model such as layers, attention heads, or specific weights) can be used to uncover necessary language knowledge within the models for successful task performance, generating hypotheses about language representations. Additionally, public release of the datasets used for training optimization will enable researchers to directly test the causal effect of input characteristics (Frank, 2023). Models can serve in controlled experiments to isolate pretraining inputs that enable effective disambiguation, offering insights into sentence-level factors that might assist children in developing word sense proficiency.

Models' performance was impaired in tasks that introduced contrastive cues (Figure 3 and 4). This suggests that this area requires further investigation, despite previous results showing that Transformers approximate adult performance in annotating word senses (Loureiro et al., 2021) or judging the semantic relatedness between word senses (Nair et al., 2020) when tested on non-contrastive sentences. Sense prototypes based on child input might have contributed to the low performance of the models in our study. In additional analyses presented in Appendix S6, we replicated all the simulations in the study using sense prototypes based on sense-tagged utterances from adult-directed speech. Specifically, we used utterances from the spoken part of the British National Corpus (BNC Consortium, 2007). We found that adult-based and child-based models produced similar percentages of correct responses in every experiment. Further, when we related models' performance to child responses, we found that child-based prototypes more closely aligned models with child performance in coherent tasks (Rabagliati et al., 2013; Study 1), but no difference was found at capturing child responses between models using child and adult sense prototypes in contrastive tasks (Rabagliati et al., 2013; Study 2; Cabiddu et al., 2022b). Overall, these supplemental results indicate that the low model performance at contrastive tasks was not due to a lack of richer linguistic cues that adult utterances might contain. The fact that the models performed poorly in tasks involving contrastive cues, even when the sense prototypes were derived from adult-directed speech, stands in contrast to the many linguistic feats of large language models (e.g., Gammelgaard et al., 2023).

Given that previous studies have not used contrastive tasks, one possibility is that such tasks might simply be difficult for models. Few models were sensitive to contrastive cues (Figure 3 and 4), indicating that at least some information about top-down structures might be captured from sentence context via distributional learning. However, overall models' performance was lower than children's. This occurs even if the task proposed to children might be more challenging than what the models faced. In fact, the models were only required to disambiguate between two alternative senses of each target word. However, other potential senses of a target word exist in dictionaries and may be known to children (e.g., for "band", not just "elastic band" and "music band", but also a “band” of bad weather). We would expect the models' ability to distinguish between word senses to deteriorate when considering a wider array of alternative senses. This is supported by Loureiro et al. (2019), which demonstrated that collapsing some of the senses in WordNet, that might not be distinguished by adults, improved Transformers' performance in word sense disambiguation.

Difficulties in approximating child knowledge could be due to the fact that children's representations of top-down structures are not only based on sentence context but also include real-world knowledge, which would need to be integrated into neural systems and could lead to abstractions more akin to human cognition (e.g., Pavlick, 2023). Specifically, while language models are capable of forming knowledge about direct word associations (bottom-up knowledge) and syntactic and semantic structures (top-down knowledge), it is crucial to acknowledge that the models' top-down generalisations about language patterns—though reflective of a form of understanding or knowledge—remain purely derived from textual patterns. For example, the models may solve experimental tasks (e.g., "Sophia listened to some music. Then, she twisted a band") by leveraging indirect associations between words—such as "twist" being associated with "bend" and "pull"—or by linking verbs to various objects (e.g., "twist" with "scarf" or "knob"), using these patterns as proxies for top-down inferences. This process enables language models to abstract semantic properties from the verbs and apply these properties to new contexts or objects that they have not explicitly encountered in their training data. The model's reliance on indirect associations to infer word meanings or predict plausible word combinations exemplifies a form of semantic generalisation. This simulates top-down processing by using the extensive network of associations encoded within their training data, thereby enabling application of these patterns to novel linguistic contexts. However, it remains an open question whether top-down abstractions based only on language patterns can approximate the generalisations that emerge from grounded representations (e.g., Pavlick, 2023, for a discussion on this topic). The challenges faced by large language models in word sense disambiguation, as highlighted in our current study, could provide valuable insights into whether grounded representations are necessary to accurately model human language processing.

For example, when modelling word acquisition trajectories, Transformers are not influenced by grounded sensorimotor, social, and cognitive factors (e.g., noun concreteness), but rely on surface features (e.g., word frequency) to a greater extent than children (Chang & Bergen, 2021). We speculate that this lack of grounded knowledge might also explain the fact that the models performed worse at disambiguating prior contexts than current contexts (Figure 2). Current contexts contained words that might appear closer to target words in naturalistic language, becoming easier to track by a distributional learner. This difficulty might not exist for children who can use their real-world knowledge for semantically-related (but distant) words (e.g., in *Dora looked in her drawer. The band was cool,* a child can infer that entities stored in a drawer are usually objects). Indeed, word acquisition trajectories can probably be better captured by neural models that process a richer multimodal signal comprising auditory features, communicative intentions, and perceptual information about word referents (e.g., Frank et al., 2009; Nikolaus & Fourtassi, 2021; Nyamapfene & Ahmad, 2007). Future work should focus on modelling child multimodal processing, currently limited by the scarcity of naturalistic multimodal corpora (e.g., Nikolaus et al., 2022).

Integrating multimodal input could also be potentially beneficial for investigating the models' performance with words varying in concreteness (e.g., concrete nouns vs more abstract verbs), which was not considered in our simulations but could be intriguing given the role of concreteness in early vocabulary learning (e.g., Braginsky et al., 2019). For instance, abstract nouns or verbs might depend more heavily on linguistic context for disambiguation, whereas concrete nouns might rely more on multimodal contexts (e.g., Sakreida et al., 2013). Highlighting this distinction could be valuable for future research, suggesting that Transformers trained on text might demonstrate superior performance with abstract words. This potential difference warrants further investigation to better understand how varying contexts influence word disambiguation across different word types.

Moreover, examining the distinction between concrete and abstract word senses could further elucidate the implications of basing model sense prototypes on child-directed or adult-directed sentences. For instance, since child-directed input often features more redundancy and a concrete vocabulary (Saxton, 2009) compared to adult-directed input, this might result in the formation of sense prototypes that better facilitate the disambiguation of concrete nouns like those used in our study. In other words, similarly to how child-directed sense prototypes may lead to a dominant sense bias typical of child-directed speech (Appendix S6), one should also find that child-directed sense prototypes lead to a bias toward concrete nouns.

Enriching models’ input would allow researchers to test if acquiring multimodal knowledge suffices to capture sensitivity to top-down structures, or whether one would need to integrate domain-specific constraints in line with nativist approaches (e.g., Pinker, 1989; Thornton, 2012) or more domain-general innate biases (e.g., Perfors et al., 2011). For instance, a development of our work might involve investigating whether a purely distributional learner that can process visual object referents is able to bootstrap certain elements of sentence structure that are posited to be innate by alternative theories of language development. For example, when a word typically used as a verb (e.g., “eat”) is presented in a noun context (e.g., “an eat”), 20-month-old infants more readily associate the word with a novel animal. Conversely, when a noun is strongly linked to a specific referent (e.g., “dog”), infants struggle to apply it to a different novel animal (Dautriche et al., 2018). This phenomenon indicates that employing different syntactic categories facilitates the extension of a word's meaning to encompass new referents. Given this evidence, one could examine whether a purely distributional learner, trained on input mirroring the quantity and quality available to 20-month-old infants exhibits similar facilitation from syntactic categories on word sense extension. Such empirical evidence would challenge the idea, proposed by universal grammar theories, that syntactic categories are innate rather than learned through language interaction (e.g., Valian et al., 2009).

Additionally, our method of assessing word sense disambiguation in large language models and humans could be used to evaluate approaches that view learning as an embodied and situated phenomenon. Indeed, the formation of semantic representations of words is not uniquely based on the statistics of word co-occurrences in language (the language-based distributional hypothesis). Properties of words related to the extralinguistic environment (e.g., physical properties) also play a crucial role in shaping semantic representations (the experiential hypothesis). Examining the capabilities of a distributional learner that relies exclusively on language co-occurrence statistics to capture word semantic representations can shed light on the importance of considering the real-world experiences of children. This approach can help determine how these two sources of information—linguistic and experiential—contribute independently or together to children's learning (e.g., Andrews et al., 2009). To this end, research involving language-based large language models can be expanded to also consider the combined influence of visual aspects (Lu et al., 2019; Qi et al., 2020; Sun et al., 2019; Zhuang et al., 2023).

Finally, the language that children are exposed to is often displaced, meaning caregivers frequently discuss word referents that are not present in the immediate environment (Tomasello & Kruger, 1992). Despite this, children might still leverage extralinguistic cues, such as iconicity (e.g., a caregiver mimicking the action of swinging a bat to clarify its meaning in conversation), in line with the language-as-situated hypothesis (Murgiano et al., 2021). Therefore, exploring the extent to which child semantic representations can be derived from both the linguistic and physical contexts in which children learn can reveal whether it is necessary to incorporate additional aspects of the communicative context, such as iconic cues, into our understanding of child word meaning representation.

Conclusion - What Large Language Models (LLMs) can('t) tell us about child language acquisition

We have begun to examine the capabilities and limitations of Transformer models for studying early word sense disambiguation. We have demonstrated that, as efficient distributional learners processing raw language input, large language models can be used to provide proof of principles concerning the extent to which usage-based learning can contribute to the acquisition of semantic representations at the word level. Importantly, it is this proficiency that highlights an interesting contrast: We have found that, although large language models excel at numerous language understanding and production tasks, they show significant limitations in their use of top-down cues for sense disambiguation. This results in their performance falling short compared to that of young children under certain disambiguation conditions. This finding serves as a crucial hint that an approach centered on providing more distributional linguistic cues might not be the most effective solution. Rather, it underscores the importance of either making models sensitive to additional multimodal cues or integrating specific constraints or biases into the models. This additional knowledge could potentially enable them to bridge the performance gap and align more closely to child learners.

Furthermore, in our tasks requiring the use of sentence context for word-level disambiguation, large language models have allowed us to avoid having to equip the models with syntactic and semantic knowledge at the sentence level (using external resources to pre-process the input) to ultimately perform word disambiguation. This would have required making assumptions about what knowledge the learner possesses at a certain point in development, which can come with benefits but also complications stemming from confounding effects caused by the assumptions made by the modeler.

Finally, we showed that an evaluation approach that leverages sense-annotated corpora can sensibly be used to examine the developmental plausibility of sense representations in large language models. Currently, limitations concerning model pretraining do not allow researchers to determine the impact of child language input on models' performance. However, we have seen that even the simple use of sense prototypes based on child input produced a partial alignment to child processing. This presents the prospect of combining corpus analyses of models' input with experimental simulations to elucidate the dynamics between the contribution of input characteristics and the nature of the learner's representational system.

References

Abbot-Smith, K., & Tomasello, M. (2006). *Exemplar-learning and schematization in a usage-based account of syntactic acquisition. 23*(3), 275–290. https://doi.org/10.1515/TLR.2006.011

Alishahi, A., & Stevenson, S. (2013). Gradual Acquisition of Verb Selectional Preferences in a Bayesian Model. In A. Villavicencio, T. Poibeau, A. Korhonen, & A. Alishahi (Eds.), *Cognitive Aspects of Computational Language Acquisition* (pp. 297–316). Springer. https://doi.org/10.1007/978-3-642-31863-4_11

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

Ambridge, B. (2020). Abstractions made of exemplars or 'You're all right, and I've changed my mind': Response to commentators. *First Language, 40*(5–6), 640–659. https://doi.org/10.1177/0142723720949723

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

Andreu, L., Sanz-Torrent, M., & Trueswell, J. C. (2013). Anticipatory sentence processing in children with specific language impairment: Evidence from eye movements during listening. *Applied Psycholinguistics, 34*(1), 5–44. https://doi.org/10.1017/S0142716411000592

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463–498. https://doi.org/10.1037/a0016261

BNC Consortium. (2007). *British National Corpus, XML edition.* https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and Variability in Children's Word Learning Across Languages. *Open Mind : Discoveries in Cognitive Science, 3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Bybee, J. (2010). *Language, Usage and Cognition.* Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022a). ChiSense-12: An English Sense-Annotated Child-Directed Speech Corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5198–5205. https://aclanthology.org/2022.lrec-1.557

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022b). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44). https://escholarship.org/uc/item/9kh29212

Chang, T. A., & Bergen, B. K. (2021). *Word Acquisition in Neural Language Models* (arXiv:2110.02406). arXiv. https://doi.org/10.48550/arXiv.2110.02406

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways* (arXiv:2204.02311). arXiv. https://doi.org/10.48550/arXiv.2204.02311

Dautriche, I., Fibla, L., Fievet, A.-C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology, 104,* 83–105. https://doi.org/10.1016/j.cogpsych.2018.04.001

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Duffy, S. A., Kambe, G., & Rayner, K. (2001). The effect of prior disambiguating context on the comprehension of ambiguous words: Evidence from eye movements. In *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 27–43). American Psychological Association. https://doi.org/10.1037/10459-002

Frank, M. C. (2023). Openly accessible LLMs can help us to understand human cognition. *Nature Human Behaviour, 7*(11), Article 11. https://doi.org/10.1038/s41562-023-01732-4

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*(5), 578–585. https://doi.org/10.1111/j.1467-9280.2009.02335.x

Gammelgaard, M. L., Christiansen, J. G., & Søgaard, A. (2023). *Large language models converge toward human-like concept organization* (arXiv:2308.15047). arXiv. https://doi.org/10.48550/arXiv.2308.15047

Haber, J., & Poesio, M. (2020). Word Sense Distance in Human Similarity Judgements and Contextualised Word Embeddings. *Proceedings of the Probability and Meaning Conference (PaM 2020),* 128–145. https://aclanthology.org/2020.pam-1.17

Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid Linguistic Ambiguity Resolution in Young Children with Autism Spectrum Disorder: Eye Tracking Evidence for the Limits of Weak Central Coherence. *Autism Research, 8*(6), 717–726. https://doi.org/10.1002/aur.1487

Hewitt, J., & Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138. https://doi.org/10.18653/v1/N19-1419

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). *Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training* (p. 2022.10.04.510681). bioRxiv. https://doi.org/10.1101/2022.10.04.510681

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. https://doi.org/10.18653/v1/2021.conll-1.49

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Khanna, M. M., & Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution. *Quarterly Journal of Experimental Psychology, 63*(1), 160–193. https://doi.org/10.1080/17470210902866664

Kidd, E., & Bavin, E. L. (2005). Lexical and referential cues to sentence interpretation: An investigation of children's interpretations of ambiguous sentences. *Journal of Child Language, 32*(4), 855–876. https://doi.org/10.1017/S0305000905007051

Laverghetta Jr, A., & Licato, J. (2021). Modeling Age of Acquisition Norms Using Transformer Networks. *The International FLAIRS Conference Proceedings, 34*. https://doi.org/10.32473/flairs.v34i1.128334

Loureiro, D., Jorge, A. M., & Camacho-Collados, J. (2022). LMMS Reloaded: Transformer-based Sense Embeddings for Disambiguation and Beyond. *Artificial Intelligence, 305*, 103661. https://doi.org/10.1016/j.artint.2022.103661

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). *Analysis and Evaluation of Language Models for Word Sense Disambiguation* (arXiv:2008.11608). arXiv. https://doi.org/10.48550/arXiv.2008.11608

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural*

Information Processing Systems, *32*. https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed* (pp. xi, 366). Lawrence Erlbaum Associates Publishers.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(4), 843–847. https://doi.org/10.1037/a0029284

Mani, N., Daum, M. M., & Huettig, F. (2016). "Proactive" in many ways: Developmental evidence for a dynamic pluralistic approach to prediction. *Quarterly Journal of Experimental Psychology*, *69*(11), 2189–2201. https://doi.org/10.1080/17470218.2015.1111395

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*, 1–51. https://doi.org/10.1037/rev0000126

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 51–61. https://doi.org/10.18653/v1/K16-1006

Meylan, S. C., Mankewitz, J., Floyd, S., Rabagliati, H., & Srinivasan, M. (2021). Quantifying Lexical Ambiguity in Speech To and From English-Learning Children. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/1pq031fn

Murgiano, M., Motamedi, Y., & Vigliocco, G. (2021). Situating Language in the Real-World: The Role of Multimodal Iconicity and Indexicality. *Journal of Cognition*, *4*(1), 38. https://doi.org/10.5334/joc.113

Nair, S., Srinivasan, M., & Meylan, S. (2020). Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge. In M. Zock, E. Chersoni, A. Lenci, & E. Santus (Eds.), *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon* (pp. 129–141). Association for Computational Linguistics. https://aclanthology.org/2020.cogalex-1.16

Nikolaus, M., & Fourtassi, A. (2021). Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 200–210. https://doi.org/10.18653/v1/2021.cmcl-1.24

Nikolaus, M., Alishahi, A., & Chrupała, G. (2022). Learning English with Peppa Pig. *Transactions of the Association for Computational Linguistics, 10*, 922–936. https://doi.org/10.1162/tacl_a_00498

Nyamapfene, A., & Ahmad, K. (2007). A Multimodal Model of Child Language Acquisition at the One-Word Stage. *2007 International Joint Conference on Neural Networks*, 783–788. https://doi.org/10.1109/IJCNN.2007.4371057

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition, 118*(3), 306–338. https://doi.org/10.1016/j.cognition.2010.11.001

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pinker, S. (1989). Learnability and Cognition. *MIT Press*. https://mitpress.mit.edu/9780262660730/learnability-and-cognition/

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., & Sacheti, A. (2020). ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv:2001.07966 [Cs]*. http://arxiv.org/abs/2001.07966

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology, 49*, 1076–1089. https://doi.org/10.1037/a0026918

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (arXiv:1910.10683). arXiv. https://doi.org/10.48550/arXiv.1910.10683

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science, 15*(2), 411–427.

https://doi.org/10.1177/1745691619885860

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science, 28*(1), 89–104. https://doi.org/10.1207/s15516709cog2801_4

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science, 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Sakreida, K., Scorolli, C., Menz, M. M., Heim, S., Borghi, A. M., & Binkofski, F. (2013). Are abstract action words embodied? An fMRI investigation at the interface between language and motor cognition. *Frontiers in Human Neuroscience, 7,* 125. https://doi.org/10.3389/fnhum.2013.00125

Saxton, M. (2009). The Inevitability of Child Directed Speech. In S. Foster-Cohen (Ed.), *Language Acquisition* (pp. 62–86). Palgrave Macmillan UK. https://doi.org/10.1057/9780230240780_4

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language, 58*(2), 574–608. https://doi.org/10.1016/j.jml.2007.08.001

Srinivasan, M., & Rabagliati, H. (2021). The Implications of Polysemy for Theories of Word Learning. *Child Development Perspectives, 15*(3), 148–153. https://doi.org/10.1111/cdep.12411

Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A Joint Model for Video and Language Representation Learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV),* 7463–7472. https://doi.org/10.1109/ICCV.2019.00756

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT Rediscovers the Classical NLP Pipeline* (arXiv:1905.05950). arXiv. https://doi.org/10.48550/arXiv.1905.05950

Thornton, R. (2012). Studies at the interface of child language and models of language acquisition. *First Language*, *32*(1–2), 281–297. https://doi.org/10.1177/0142723711403881

Tomasello, M., & Kruger, A. C. (1992). Joint attention on actions: Acquiring verbs in ostensive and non-ostensive contexts. *Journal of Child Language*, *19*(2), 311–333. https://doi.org/10.1017/S0305000900011430

Trueswell, J. C., & Gleitman, L. R. (2007). Learning to parse and its implications for language acquisition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568971.013.0039

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, *36*(4), 743–778. https://doi.org/10.1017/S0305000908009082

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Warstadt, A., & Bowman, S. R. (2022). *What Artificial Neural Networks Can Tell Us About Human Language Acquisition* (arXiv:2208.07998). arXiv. https://doi.org/10.48550/arXiv.2208.07998

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell (Eds.), *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 1–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-babylm.1

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). *Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings* (arXiv:1909.10430). arXiv. https://doi.org/10.48550/arXiv.1909.10430

Yacovone, A., Shafto, C. L., Worek, A., & Snedeker, J. (2021). Word vs. World Knowledge: A developmental shift from bottom-up lexical cues to top-down plausibility. *Cognitive Psychology*, *131*, 101442. https://doi.org/10.1016/j.cogpsych.2021.101442

Zhang, Y., Warstadt, A., Li, X., & Bowman, S. R. (2021). When Do You Need Billions of Words of Pretraining Data? *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1112–1125. https://doi.org/10.18653/v1/2021.acl-long.90

Zhuang, C., Fedorenko, E., & Andreas, J. (2023, October 20). *Visual Grounding Helps Learn Word Meanings in Low-Data Regimes*. arXiv.Org. https://arxiv.org/abs/2310.13257v1

Data, code and materials availability statement

Raw data, simulation and analysis scripts used in the study can be found on the GitHub project repository https://doi.org/10.5281/zenodo.8200803. The ChiSense-12 corpus can be downloaded at https://gitlab.com/francescocabiddu/chisense-12. The CHILDES database is accessible at https://childes.talkbank.org/. The British National Corpus can be downloaded at https://llds.ling-phil.ox.ac.uk/llds/xmlui/handle/20.500.14106/2554.

Ethics statement

Ethics approval was not required as the study used previously-collected publicly-available data from the CHILDES database (MacWhinney, 2000).

Authorship and Contributorship Statement

Francesco Cabiddu: conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; software; writing – original draft preparation; writing – review & editing; funding acquisition. **Mitja Nikolaus:** conceptualization; formal analysis; methodology; writing – review & editing. **Abdellah Fourtassi**: conceptualization; formal analysis; methodology; writing – review & editing.

All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We would like to express our gratitude to Alex Warstadt for his valuable comments during the initial stages of this project, and to Abhishek Agrawal for offering helpful suggestions on the manuscript. We also extend our thanks to the anonymous reviewers for their insightful feedback. The research presented in this article has been generously supported by a Study Visit Grant from the Experimental Psychology Society (UK).

Appendix S1: Model Families

We provide a description of the model families included in the study, and details about models' configurations varying in model size and pretraining size (Table S1.1). Transformer models were downloaded using the Huggingface Transformers Python library (Wolf et al. 2020), apart from the model BabyBERTa (Huebner et al. 2021) whose pretrained weights were downloaded directly from its GitHub project page (https://github.com/phueb/BabyBERTa, October 2022). The recurrent neural model ELMo (version 3; Peters et al. 2018) was downloaded using the TensorFlow Python library (Abadi et al. 2015).

The 13 Transformer model families used were: BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and GTP (OpenAI GPT, Radford et al. 2018; GPT-2, Radford et al. 2019). For each of these three families we included their distilled model versions (DistilBERT, DistilRoBERTa, and DistilGPT2; Sanh et al. 2020), and the RoBERTa family also included versions pretrained on small corpora (MiniBERTa, Warstadt et al. 2020). BabyBERTa (Huebner et al. 2021); ALBERT-v1 and ALBERT-v2 (Lan et al. 2020); DeBERTa and DeBERTA-v2 (He, Gao, et al., 2021); DeBERTa-v3 (He, Liu, et al., 2021); Transformer-XL (Dai et al., 2019); CTRL (Keskar et al., 2019); T5 (Raffel et al., 2020); XLNet (Yang et al., 2020).

A first macro distinction between families concerns their unidirectional or bidirectional way of predicting a token given its context. Unidirectional Transformers (GPT, Transformer-XL, and CTRL) are trained on predicting the next token given the (previous) left sentence context. This type of training objective is in line with prediction-based approaches of children's online sentence processing (Mani & Huettig, 2012). The remaining Transformers and ELMo are instead trained on predicting tokens by taking into account both (previous) left and (following) right contexts. This type of objective is plausible because children are not only involved in predicting upcoming input when hearing speech, but they can also revise their interpretation of ambiguous words based on following cues (e.g., Qi et al., 2020). Also, in naturalistic conversations there are cases in which children would likely attend to following sentence context to disambiguate nouns (e.g., "*Look at the bat, it's flying!*").

A second macro distinction concerns how different models track the position of tokens in a text sequence. Most models track tokens' absolute positions, essentially encoding sentence word order which is required for learning syntax (e.g., distinguishing between "*The dog chased the boy*" and "*The boy chased the dog*"). Additionally, some models implement mechanisms that track both absolute and relative positions of tokens (DeBERTa, DeBERTa-v2, DeBERTa-v3) or only relative positions (Transformer-XL, T5, XLNet). Tracking relative positions means tracking the relative distance between pairs of tokens in a sequence, which translates into weighting more the words that appear closer to a target word (e.g., the contribution of "*deep*" for the vector representation of "*learning*" is higher if the two appear one next to the other, compared to when they appear in different sentences). Tracking relative positions can be considered a proxy of children's sentence local parsing (e.g., Gertner & Fisher, 2012).

BERT is a bidirectional Transformer trained on predicting tokens that are masked at random during the preprocessing of the input, with some sentences seen multiple times with the same masked tokens (i.e., static masking). It is also pretrained on predicting whether a sentence follows another in the input (next sentence prediction), with the aim of capturing relations between sentences that can be useful in Question Answering and Natural Language Inference tasks. The model is pretrained on the BookCorpus (Zhu et al., 2015) and English Wikipedia.

RoBERTa is a modification of BERT that is trained without the next sentence prediction objective, which investigations found to be not effective for improving performance in downstream tasks (e.g., Liu et al. 2019; Yang et al., 2020). It is trained by receiving larger batches of examples at every weight updating iteration. It is also trained on a larger corpus than BERT, additionally including English news articles, web content, and stories. The model also uses dynamic masking, which masks different tokens every time the same sentence is fed to the model. Its scaled-down version, MiniBERTa, is pretrained on similar input (BookCorpus and Wikipedia) but on a much smaller scale (see Table S1.1), with the configuration pretrained on the smallest corpus (1M tokens) also reduced in model size.

GPT models are unidirectional Transformers trained on a language modeling objective, namely sampling text from the input dataset and asking the model to predict the next token. OpenAI GPT was pretrained on the BookCorpus, and subsequently fine-tuned with a series of supervised language understanding tasks. GPT-2 was instead pretrained on a larger corpus of web content, with no supervised fine-tuning.

Distilled models are compressed and faster versions of the above models, based on the same architectures but with reduced number of layers. They undergo training that specifically tries to reproduce the behavior of the (parent) larger model.

BabyBERTa is a scaled-down version of RoBERTa with some key differences. It is significantly reduced in size (15x fewer parameters). It modifies the masked word prediction objective: In BERT and RoBERTa, 10% of the tokens selected for masking are left unmasked; BabyBERTa never allows unmasking. It is also pretrained on much smaller (6000x fewer tokens) and qualitatively different corpora, either separately on transcribed child-directed speech, written child-directed news articles, a small portion of Wikipedia, or a combination of the three.

ALBERT-v1 is a light version of BERT that was created with the main goal of reducing the computational costs derived from using a large number of parameters. ALBERT-v1 uses two techniques (factorization of parameters, and sharing all parameters across model layers) which significantly reduce the number of parameters without significant drops in performance in downstream tasks. Additionally, ALBERT-v1 modifies the next sentence prediction objective performing sentence order prediction instead. A key difference between the two objectives is that in next sentence prediction, the model is provided with positive examples of pairs of consecutive sentences coming from the same document, and negative examples with the second sentence of the pair swapped with one coming from a different document. The inefficiency of this task comes from the fact that negative examples contain sentences coming from different documents, which likely contain text about different topics. This results in the model being able to easily learn from negative examples by just noticing differences in word occurrences (i.e., semantically different words are used when sentences refer to different topics), focusing less on the more important aspect of discourse coherence between the two sentences. Therefore, with the new sentence order prediction objective, negative examples comprise sentence pairs coming from the same document, just swapped in order. This forces the model to focus on the coherence of one sentence following the other. This new objective significantly improved performance in downstream tasks compared to BERT. ALBERT-v1 is pretrained on the same datasets used for BERT.

ALBERT-v2 is a modification of ALBERT-v1 that improves performance at downstream tasks by using a different training regime (higher training steps and time) and by removing dropout, which is normally used to avoid that a model overfits the training dataset.

DeBERTa is a modification of RoBERTa which improves performance in downstream tasks by using mechanisms of disentangled attention and enhanced mask decoding, which essentially allow the model to integrate both absolute and relative token positions in its vector representations. DeBERTa is trained on the same corpora used for RoBERTa but excluding English news articles.

DeBERTa-v2 is an optimized version of DeBERTa, which uses a larger vocabulary, larger pretraining dataset, and larger model sizes. It shares parameters that track sentence content and relative positions to reduce model complexity. It also integrates an additional layer in the model to better learn knowledge about subword n-grams, with the aim of more precisely tracking sentence local dependences. DeBERTa-v2 is pretrained on the same RoBERTa corpora.

DeBERTa-v3 is a modification of DeBERTa-v2 that replaces the masked word prediction objective with a replaced token detection objective, which instead of randomly masking tokens during training it replaces them with plausible (but incorrect) ones. This changes the objective of the model from having to generate plausible tokens to having to discriminate between two semantically related tokens to decide which is the appropriate one in a sentence. DeBERTa-v3 is pretrained on the same RoBERTa corpora.

Transformer-XL is a unidirectional model that uses a language modeling objective as GPT. Transformer-XL introduces a recurrence mechanism in the Transformer architecture. Usually, Transformers process input in the form of text segments of a maximum length, which results in the impossibility of modelling dependencies across segments (which are treated independently). Transformer-XL uses a mechanism that recycles hidden states of previous segments and uses them as extended context for newly processed ones. Additionally, the model introduces a new mechanism that can keep track of the relative position of tokens across different segments. Recurrence and relative positional encoding allow Transformer-XL to track short-range and long-range text dependencies, which can be used to generate very long and relatively coherent articles. The model is pretrained on a small dataset of Wikipedia articles (Merity et al., 2016).

CTRL is another unidirectional Transformer that uses a language modeling objective. However, in this model the objective is modified so that the model predicts the next token of a sequence also taking into account specific codes present in the structure of the training data. These codes give information such as the specific domain of the text being processed (e.g., Wikipedia, Books), the specific style used (e.g., Horror, Science), or the specific tasks being processed (e.g., question answering, translation). These codes are extracted directly from structural components of the training data, and ultimately allow the model to better constrain its text generation process. CTRL is pretrained on a large corpus from Wikipedia, web content including news articles and Amazon reviews, translation datasets from European parliament and United Nations proceedings, and various question-answering datasets.

T5 is a bidirectional Transformer that uses an Encoder and a Decoder architecture similar to the original Transformer (Vaswani et al., 2017). It is trained on a masked prediction objective similar to BERT, representing both single (as in BERT) and sequences of tokens in the Encoder and using learned representations to generate text in the Decoder. In our study, we only used the Encoder part of the model. The model also uses a mechanism of relative positional encoding. The model is trained on the largest corpus considered in our study, which comprises scraped content from the web.

XLNet is a bidirectional Transformer that modifies the BERT training objective using a permutation modeling objective. In BERT, masked tokens within a text sequence are predicted independently from one another. In XLNet the prediction also takes into account the relations between masked tokens. Additionally, XLNet only uses a mechanism of relative positional encoding. The model is trained on the same corpora used for BERT, with the addition of various corpora of web content and news articles.

ELMo is a bidirectional recurrent neural network model. Its mechanism of recurrence allows to link current word representations to previous ones in a text sequence. This is achieved by processing input at different timesteps, and feeding the output of previous timesteps to the current one. The recurrence mechanism leads to contextualized representations that also encode information about word order. In ELMo, the input sequence is fed to the model from left to right, and again from right to left. The two output vectors are then combined to obtain a bidirectional representation. The model is trained on a corpus of News Crawl data (Chelba et al., 2014).

Table S1.1. ***Models included in the study, by pretraining size (gigabytes of text), model size (million parameters), and family type.***

Model	**Model size**	**Pretraining size**	**Family**
distilbert-base-uncased	66	16	bert
bert-base-uncased	110	16	bert
bert-large-uncased	340	16	bert
bert-large-uncased-whole-word-masking	340	16	bert
distilroberta-base	82	40	roberta
roberta-base	125	160	roberta
roberta-large	355	160	roberta
roberta-med-small-1M-2	45	0.005	roberta
roberta-base-10M-2	125	0.05	roberta
roberta-base-100M-2	125	0.5	roberta
roberta-base-1B-3	125	5	roberta
albert-base-v1	11	16	albert-v1
albert-large-v1	17	16	albert-v1
albert-xlarge-v1	58	16	albert-v1
albert-xxlarge-v1	223	16	albert-v1
albert-base-v2	11	16	albert-v2
albert-large-v2	17	16	albert-v2
albert-xlarge-v2	58	16	albert-v2
albert-xxlarge-v2	223	16	albert-v2
deberta-base	140	80	deberta
deberta-large	400	80	deberta
deberta-xlarge	750	80	deberta
deberta-v2-xlarge	900	160	deberta-v2

Table S1.1 (continued).

Model	**Model size**	**Pretraining size**	**Family**
deberta-v2-xxlarge	1500	160	deberta-v2
deberta-v3-small	141	160	deberta-v3
deberta-v3-base	184	160	deberta-v3
deberta-v3-large	434	160	deberta-v3
babyberta-ao-childes	8	0.02	babyberta
babyberta-ao-newsela	8	0.02	babyberta
babyberta-wikipedia-1	8	0.02	babyberta
babyberta-ao-childes-ao-newsela-wikipedia-1	8	0.06	babyberta
distilgpt2	82	40	gpt
openai-gpt	116	3	gpt
gpt2	124	40	gpt
gpt2-medium	355	40	gpt
gpt2-large	774	40	gpt
gpt2-xl	1558	40	gpt
transfo-xl-wt103	284	0.4	transfo-xl
ctrl	1630	140	ctrl
t5-small	35	806	t5
t5-base	110	806	t5
t5-large	335	806	t5
xlnet-base-cased	117	126	xlnet
xlnet-large-cased	360	126	xlnet
elmo	93	4.2	elmo

References (for Appendix S1)

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). *One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling* (arXiv:1312.3005). arXiv. https://doi.org/10.48550/arXiv.1312.3005

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context* (arXiv:1901.02860). arXiv. https://doi.org/10.48550/arXiv.1901.02860

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Gertner, Y., & Fisher, C. (2012). Predicted errors in children's early sentence comprehension. *Cognition*, *124*(1), 85–94. https://doi.org/10.1016/j.cognition.2012.03.010

He, P., Gao, J., & Chen, W. (2021). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing* (arXiv:2111.09543). arXiv. https://doi.org/10.48550/arXiv.2111.09543

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention* (arXiv:2006.03654). arXiv. https://doi.org/10.48550/arXiv.2006.03654

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. https://doi.org/10.18653/v1/2021.conll-1.49

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). *CTRL: A Conditional Transformer Language Model for Controllable Generation* (arXiv:1909.05858). arXiv. https://doi.org/10.48550/arXiv.1909.05858

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* (arXiv:1909.11942). arXiv. https://doi.org/10.48550/arXiv.1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 843–847. https://doi.org/10.1037/a0029284

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). *Pointer Sentinel Mixture Models* (arXiv:1609.07843). arXiv. https://doi.org/10.48550/arXiv.1609.07843

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Qi, Z., Love, J., Fisher, C., & Brown-Schmidt, S. (2020). Referential context and executive functioning influence children's resolution of syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(10), 1922. https://doi.org/10.1037/xlm0000886

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. https://gluebenchmark.com/leaderboard.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (arXiv:1910.10683). arXiv. http://arxiv.org/abs/1910.10683
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Warstadt, A., Zhang, Y., Li, H.-S., Liu, H., & Bowman, S. R. (2020). *Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually)* (arXiv:2010.05358). arXiv. https://doi.org/10.48550/arXiv.2010.05358

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. https://doi.org/10.48550/arXiv.1910.03771

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (arXiv:1906.08237). arXiv. https://doi.org/10.48550/arXiv.1906.08237

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books* (arXiv:1506.06724). arXiv. https://doi.org/10.48550/arXiv.1506.06724

Appendix S2: Children's Target Words and Additional Annotations

In this section, we report details about the target ambiguous words used and their corresponding child-directed sentences. We used sentences from ChiSense-12 (Cabiddu et al., 2022a), a collection 53 sense-tagged corpora of American and British English child-directed speech from the CHILDES database (MacWhinney, 2000), involving 958 target children of up to 4 years of age (59 months). We selected sentences referring to 9 of the 12 ambiguous words present in the corpus, each in their dominant and subordinate sense. The remaining 3 words (flower/flour, moose/mousse, sun/son) could not be used because they had different spelling, creating no ambiguity for models' processing. Table S2.1 provides information about the number of sentences for each sense.

Some target words in the behavioral experiments were not covered by ChiSense-12. Thus, we additionally tagged all not covered words for which 40 sentences per sense were available in the same corpora used for ChiSense-12. This resulted in tagging 4 new ambiguous words (*fish* = animal/food; *lamb* = animal/food; *turkey* = animal/food; *card* = playing card/greetings card). In total, we covered 13/24 and 4/6 target words in Rabagliati et al. (2013) experiment 1 and 2 respectively, and 9/12 words from Cabiddu et al. (2022b). The sentence test items for each experiment are available in the appendices of the two original papers (Cabiddu et al., 2022b; Rabagliati et al., 2013), and in the file *test_utterances.csv* included in the R project folder of our GitHub project. The complete sets of utterances from ChiSense-12 and the new annotated words are available in the R folder of our project.

Loureiro et al. (2021) showed that a nearest neighbor approach for computing sense prototypes is stable even when drastically reducing the number of examples for each target sense. Given that we sampled a limited number of examples for each new target sense to keep the annotation work manageable (n=40), we verified that Loureiro's findings were supported in our case. We repeated the three modeling experiments using only the 9 target words of ChiSense-12, downsampling sentences for each sense before computing sense prototypes. The procedure was repeated 10 times, each time sampling a subset (n = 40) of randomly selected sentences for each sense. The results of the three experiments (Figure S2.1, S2.2, and S2.3) showed that performance remained stable even when using only 40 random sentences per sense, which justified the inclusion of the newly annotated words in our study.

Specifically, all three experiments yielded high correlations between the mean performance of each model across random samples and the performance using the full set of utterances: Experiment 1 (Rabagliati et al., 2013) r_s = .95; Experiment 2 (Rabagliati et al., 2013) r_s = .95; Experiment 1 (Cabiddu et al., 2022b) r_s = .94.

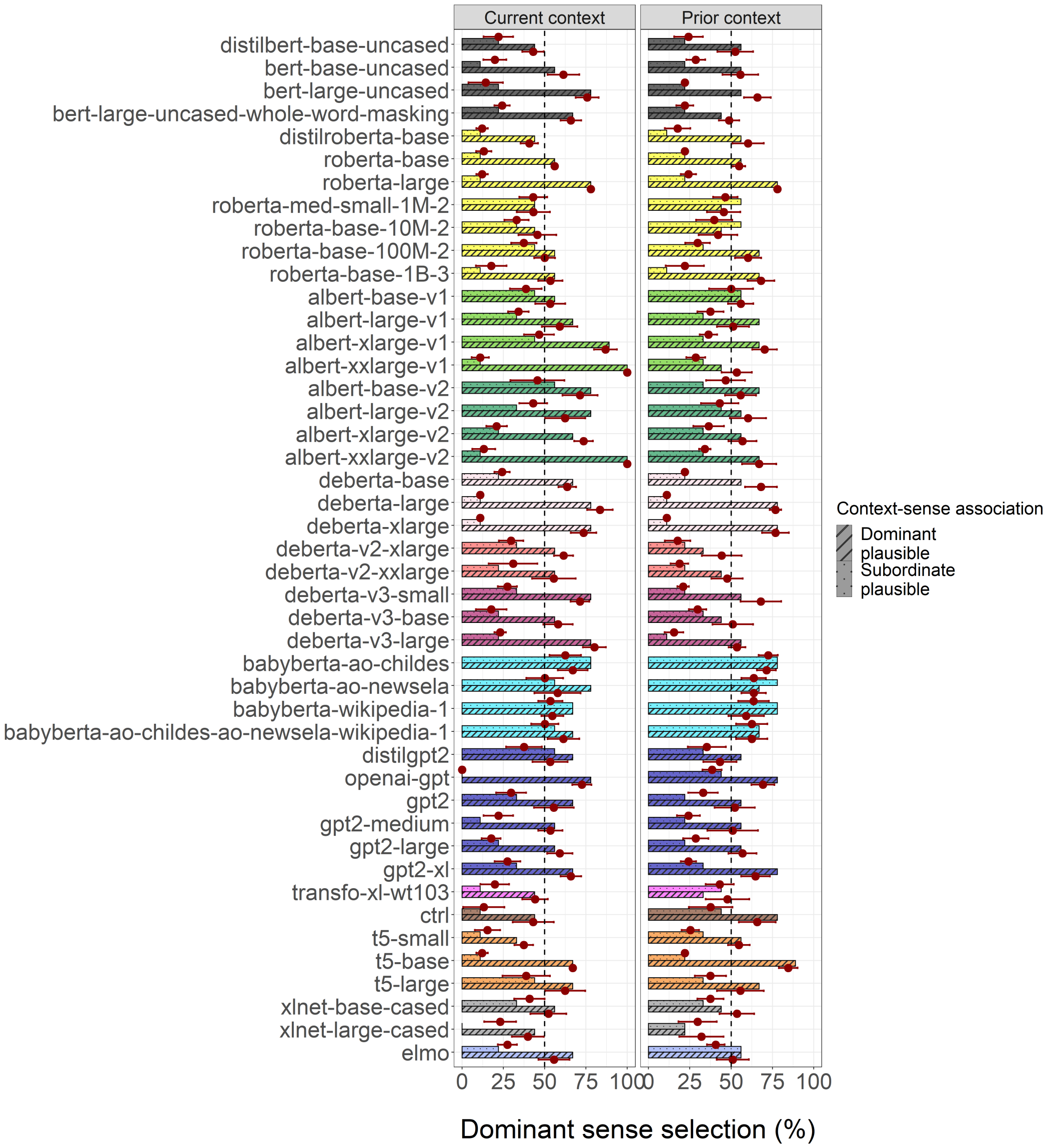

Figure S2.1. ***Percentage of dominant sense selections for experiment 1 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend), when disambiguation cues were included in current or prior context (left and right panel respectively). Colored bars indicate performance of the models when the full sample of ChiSense-12 sentences is used to compute sense prototypes (Table S2.1). Red points indicate mean performance (across 10 runs) of models for which sense prototypes were computed using 40 random sentences for each sense. Error bars indicate standard deviations.***

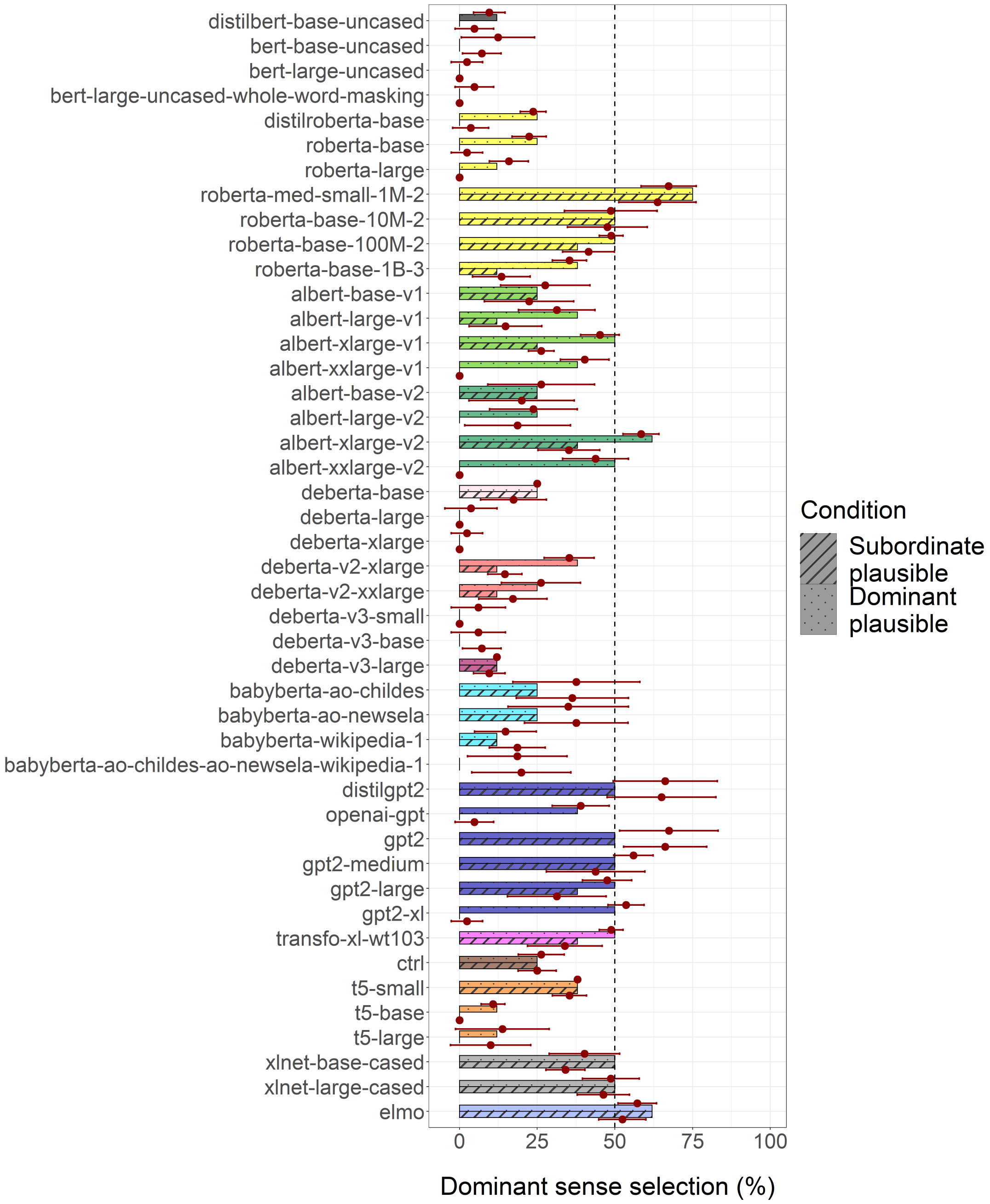

Figure S2.2. ***Percentage of dominant sense selections for experiment 2 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend). The plot shows the comparison between dominant sense selection in models with prototypes computed from the full ChiSense-12 (colored bars), and models for which prototypes were computed by downsampling ChiSense-12 to 40 random sentences per sense (points and error bars indicate mean and standard deviations across 10 runs).***

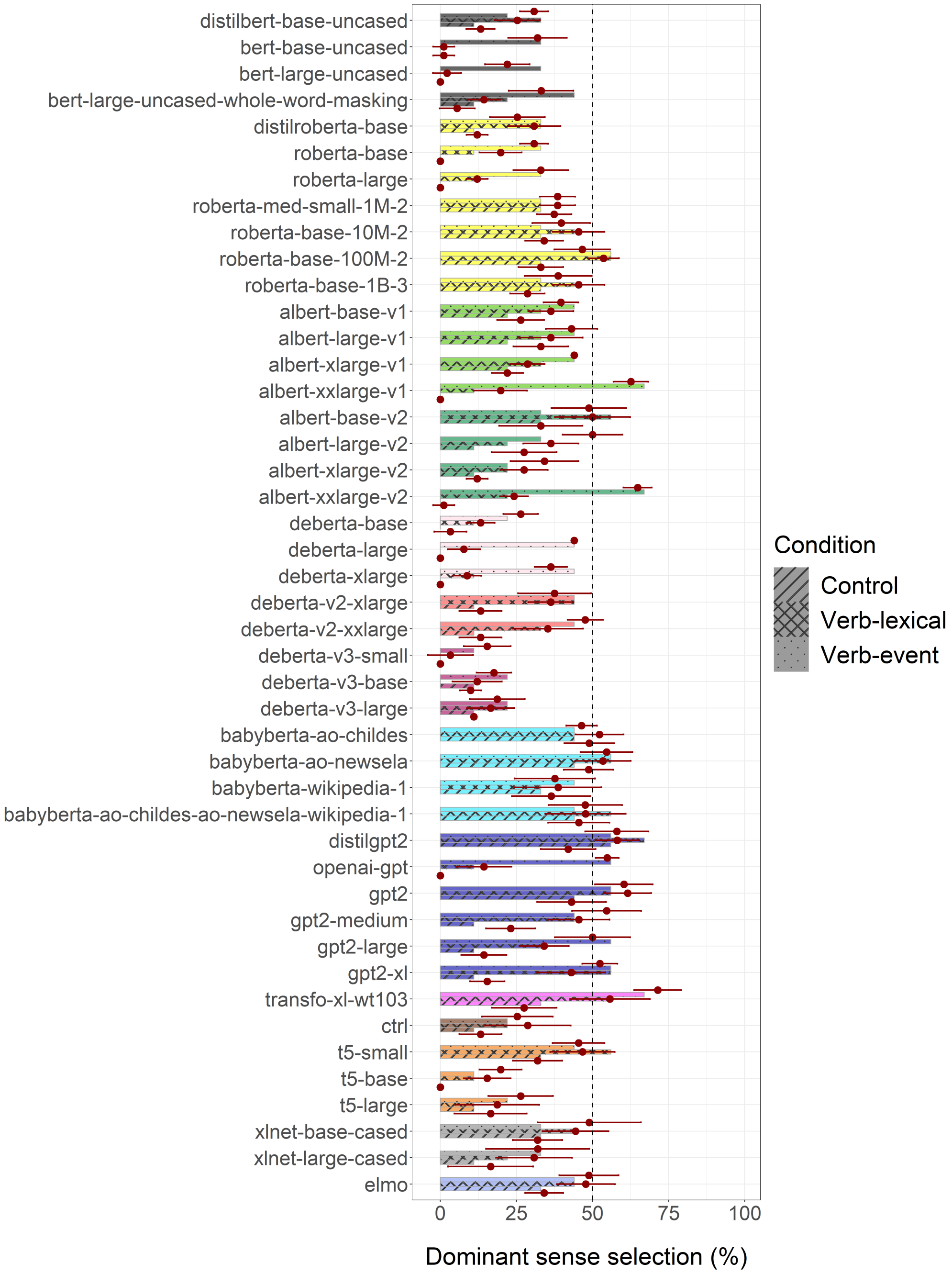

Figure S2.3. ***Percentage of dominant sense selections for Cabiddu et al. (2022b), in dominant-plausible (Verb-lexical, Verb-event) and subordinate-plausible conditions (Control). The plot shows the comparison between dominant sense selection in models with prototypes computed from the full ChiSense-12 (colored bars), and models for which prototypes were computed by downsampling ChiSense-12 to 40 random sentences per sense (points and error bars indicate mean and standard deviations across 10 runs).***

Table S2.1. ***For each target word, the table shows the raw number of utterances in which dominant (D) and subordinate (S) senses appeared, as well as the percentage of utterances in which dominant senses appeared (Dominance).***

Word (D/S)	*N* (D/S)	*Dominance*
Band (Object/Music Group)	178/58	75%
Bat (Animal/Object)	247/130	66%
Bow (Knot/Weapon)	230/27	89%
Button (Electronic/Clothing)	568/285	67%
Chicken (Animal/Food)	1463/937	61%
Glasses (Eye/Drinking)	683/620	52%
Letter (Alphabet/Mail)	1446/946	60%
Line (Geometric/Row)	471/241	66%
Nail (Finger/Tool)	460/106	81%
MEAN (*SD*)	-	69% (11%)

References (for Appendix S2)

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022a). ChiSense-12: An English Sense-Annotated Child-Directed Speech Corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5198–5205. https://aclanthology.org/2022.lrec-1.557

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022b). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44). https://escholarship.org/uc/item/9kh29212

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). *Analysis and Evaluation of Language Models for Word Sense Disambiguation* (arXiv:2008.11608). arXiv. https://doi.org/10.48550/arXiv.2008.11608

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed* (pp. xi, 366). Lawrence Erlbaum Associates Publishers.

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology, 49*, 1076–1089. https://doi.org/10.1037/a0026918

Appendix S3: Randomly Initialized Models

We modelled the three experiments (Cabiddu et al., 2022; Rabagliati et al., 2013), running base model versions 10 times using different random initializations. For a single run, the same initialization was used to create both sense prototypes and vectors of test stimuli. None of the models showed sensitivity to sentence context across experiments (Figure S3.1, S3.2, and S3.3; i.e., same percentage of dominant sense selections across conditions), suggesting that different patterns of connections among units did not influence models' performance.

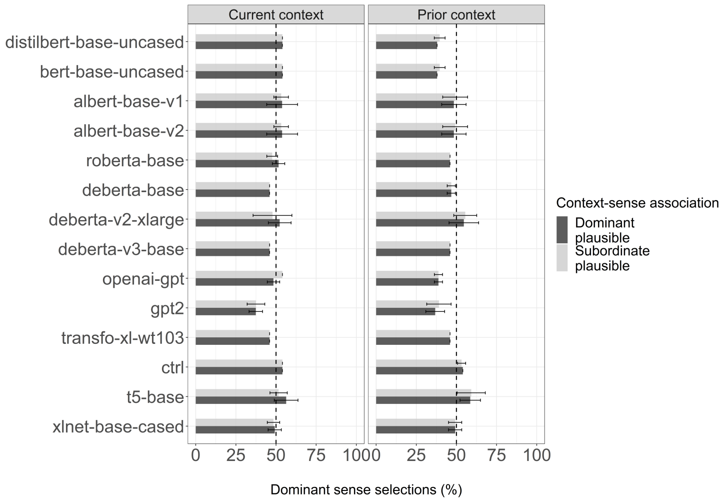

Figure S3.1. ***Mean percentage of dominant sense selections in randomly initialized models for experiment 1 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend), when disambiguation cues were included in current or prior context (left and right panel respectively). Error bars indicate standard deviations over 10 model runs.***

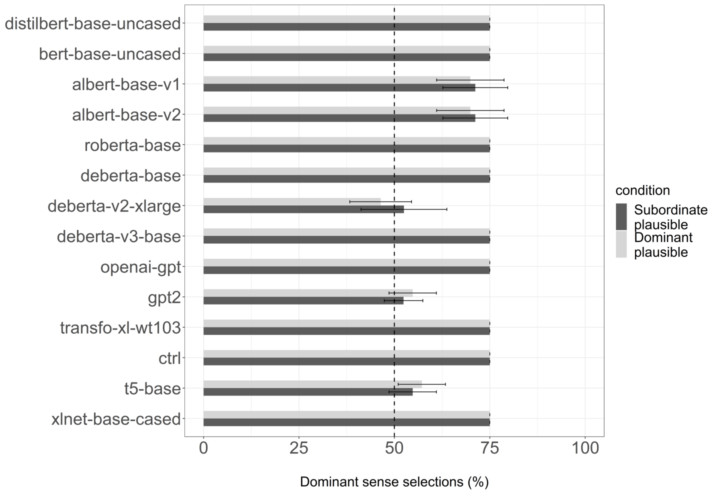

Figure S3.2. ***Mean percentage of dominant sense selections in randomly initialized models for experiment 2 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend). Error bars indicate standard deviations over 10 model runs.***

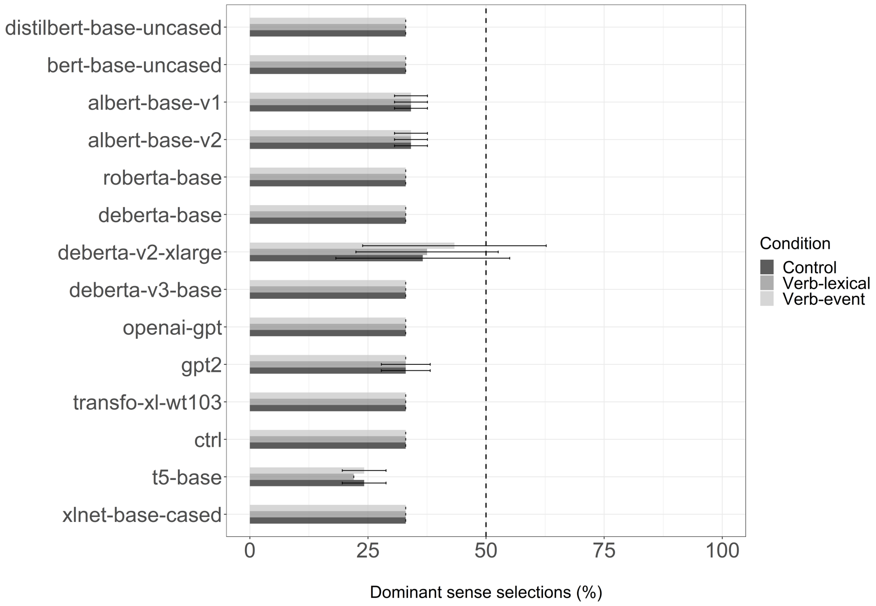

Figure S3.3. ***Mean percentage of dominant sense selections in randomly initialized models for Cabiddu et al. (2022), in dominant-plausible (Verb-lexical, Verb-event) and subordinate-plausible conditions (Control). Error bars indicate standard deviations over 10 model runs.***

References (for Appendix S3)

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/9kh29212

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, *49*, 1076–1089. https://doi.org/10.1037/a0026918

Appendix S4: Relative Difference Outcome Measure

In this section, we present the results concerning the evaluation of dominance sense preference in Transformers, using child-based prototypes. This section additionally includes plots illustrating the raw performance of each model in each of the three experiments considered. Moreover, we report the output of statistical models, where the comparison between children and models' performance is made using a measure of relative difference as the outcome (see main manuscript for details about this measure).

Dominant Bias

Table S4.1. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the percentage of dominant senses selected across conditions of Rabagliati et al. (2013) experiment 1. The predictors are log model size, log pretraining size, and their interaction. Model family was used as random effect intercept. The Null model only includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	313.52	318.94	-153.76	307.52	-	-	-
+ Model size	4	304.41	311.63	-148.20	296.41	11.11	1	**0.001**
+ Pretraining	5	293.23	302.26	-141.61	283.23	13.18	1	**0.000**
+ Interaction	6	294.46	305.30	-141.23	282.46	0.76	1	0.382

Table S4.2. ***Output of the best model selected via model comparison in Table S4.1***

	"+ Pretraining" Model Dominant sense preference		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	60.89	53.53 – 68.26	**<0.001**
Model size [log]	-1.47	-3.01 – 0.08	0.062
Pretraining size [log]	-1.53	-2.30 – -0.75	**<0.001**
Random Effects			
σ^2	25.86		
$\tau_{00\ \text{family}}$	12.47		
ICC	0.33		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.491 / 0.657		

Rabagliati et al. (2013) – Experiment 1

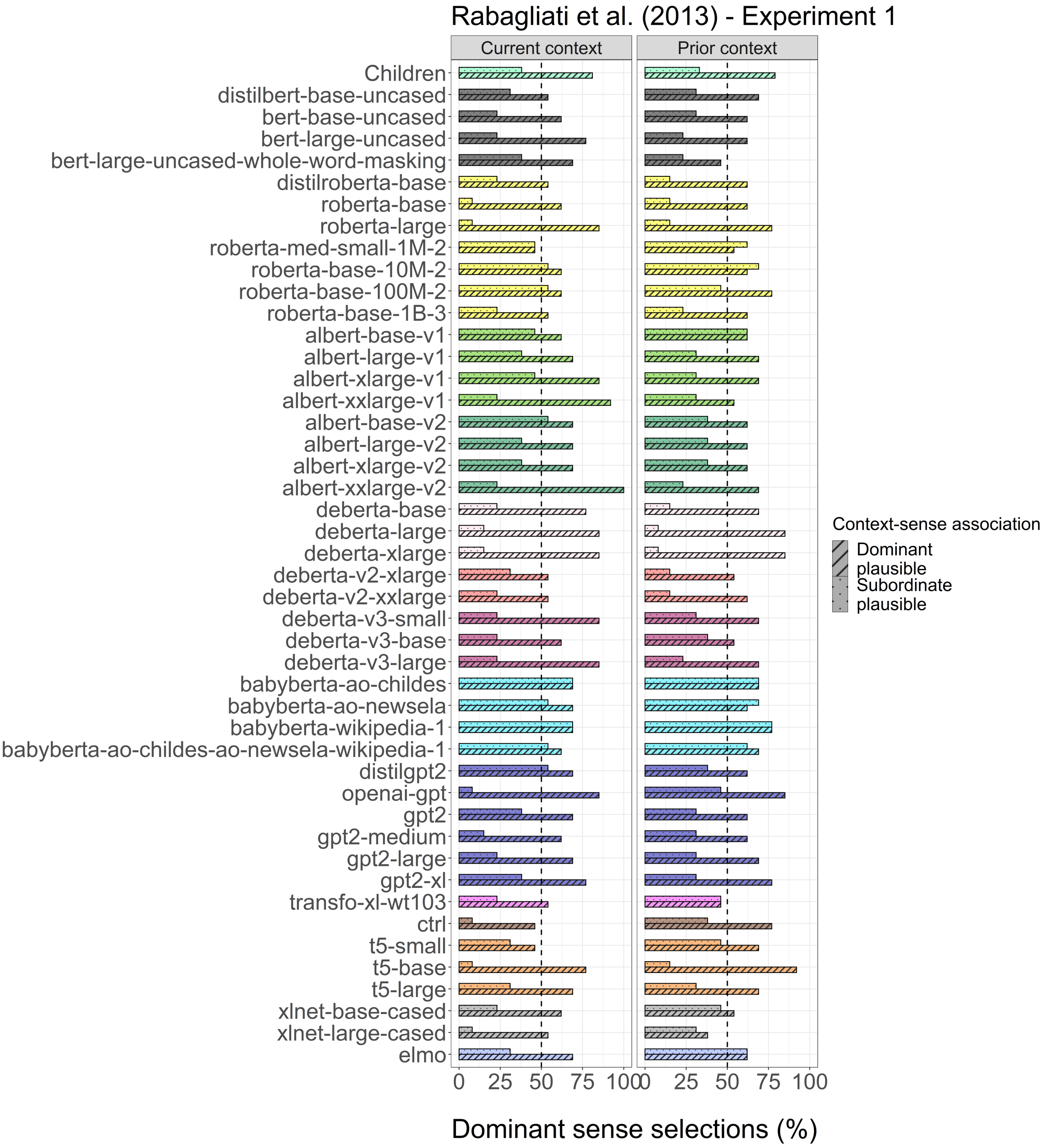

Figure S4.1. ***Percentage of dominant sense selections in models and children for experiment 1 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend), when disambiguation cues were included in current or prior context (left and right panel respectively).***

Table S4.3. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the relative difference between models and children in Rabagliati et al. (2013) experiment 1, see our main paper for more details about this outcome measure. The predictors are condition (Prior or Current context), log pretraining size, log model size , and their pairwise interactions. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	807.62	815.12	-400.81	801.62	NA	NA	NA
+ Condition	4	803.09	813.09	-397.55	795.09	6.53	1	**0.011**
+ Pretraining	5	771.98	784.47	-380.99	761.98	33.12	1	**0.000**
+ Model size	6	764.64	779.64	-376.32	752.64	9.33	1	**0.002**
+ Pretraining*Condition	7	766.34	783.84	-376.17	752.34	0.30	1	0.584
+ Size*Condition	8	768.04	788.04	-376.02	752.04	0.30	1	0.584
+ Pretraining*Model size	9	769.64	792.14	-375.82	751.64	0.40	1	0.529

Table S4.4. ***Output of the best model selected via model comparison in Table S4.3.***

	'+ Model size' model Rabagliati et al. (2013) - Experiment 1		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-43.49	-59.94 – -27.04	**<0.001**
Model size [log]	5.36	2.07 – 8.64	**0.002**
Pretraining size [log]	3.81	2.16 – 5.47	**<0.001**
Condition [Prior context]	-9.98	-16.18 – -3.78	**0.002**
Random Effects			
σ^2	218.67		
$\tau_{00\ \text{family}}$	85.81		
ICC	0.28		
N_{family}	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.510 / 0.648		

Rabagliati et al. (2013) – Experiment 2

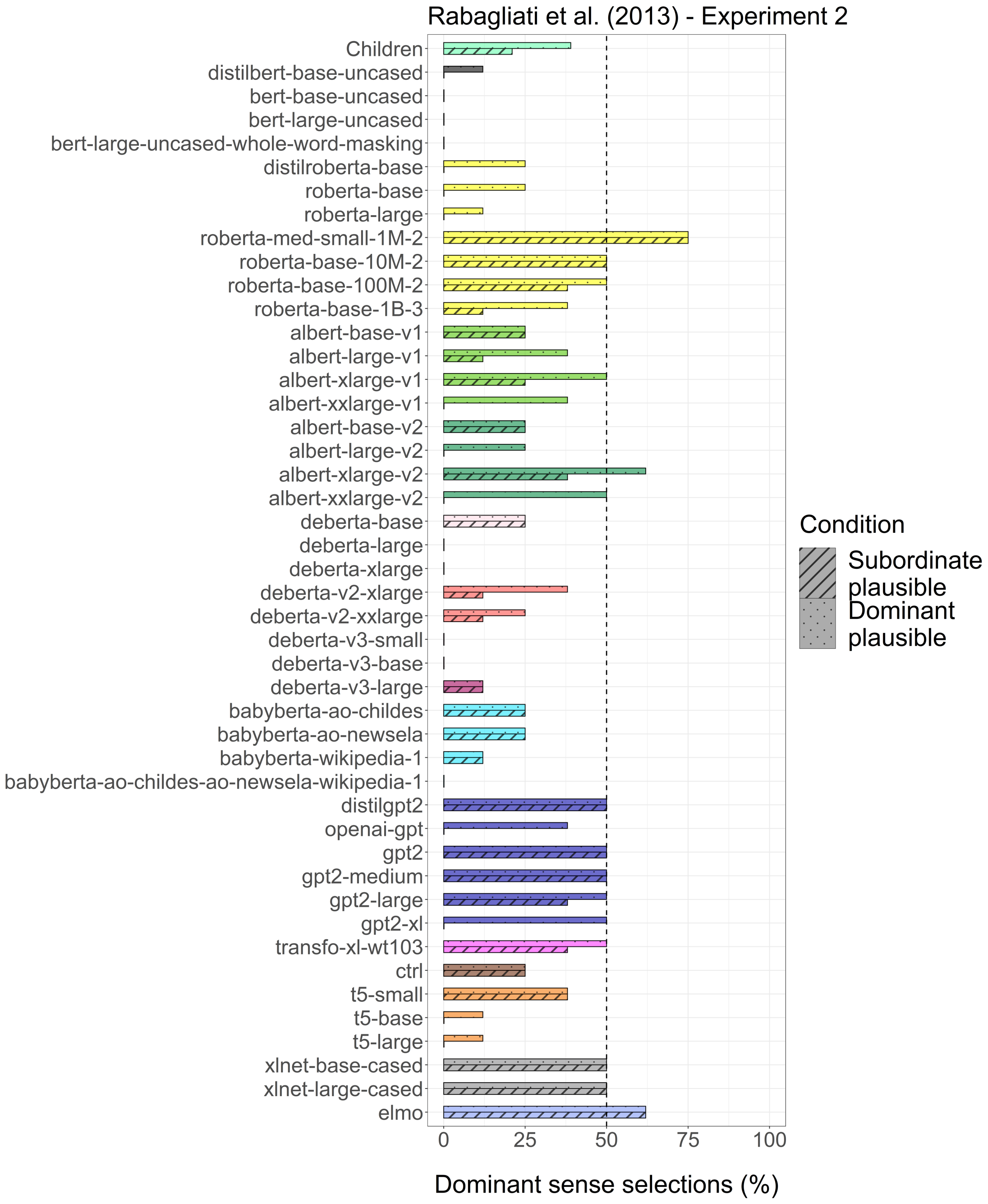

Figure S4.2. ***Percentage of dominant sense selections in models and children for experiment 2 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend).***

Table S4.5. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the relative difference between models and children in Rabagliati et al. (2013) experiment 2, see our main paper for more details about this outcome measure. The predictors are log pretraining size, log model size, and their interaction. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	371.50	376.92	-182.75	365.50	-	-	-
+ Pretraining	4	372.70	379.93	-182.35	364.70	0.80	1	0.371
+ Model size	5	372.27	381.30	-181.13	362.27	2.44	1	0.119
+ Interaction	6	373.79	384.63	-180.89	361.79	0.48	1	0.489

Table S4.6. ***Although no model surpassed the Null model in Table S4.5, below we show the output of the model including both main effects of model size and pretraining size, to appreciate size of the estimates and variance explained.***

	'+ Model size' model Rabagliati et al. (2013) - Experiment 2		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-25.51	-43.38 – -7.63	**0.006**
Model size [log]	3.37	-0.35 – 7.09	0.075
Pretraining size [log]	0.12	-1.74 – 1.98	0.895
Random Effects			
σ^2	146.43		
$\tau_{00 \text{ family}}$	79.91		
ICC	0.35		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.105 / 0.421		

Cabiddu et al. (2022)

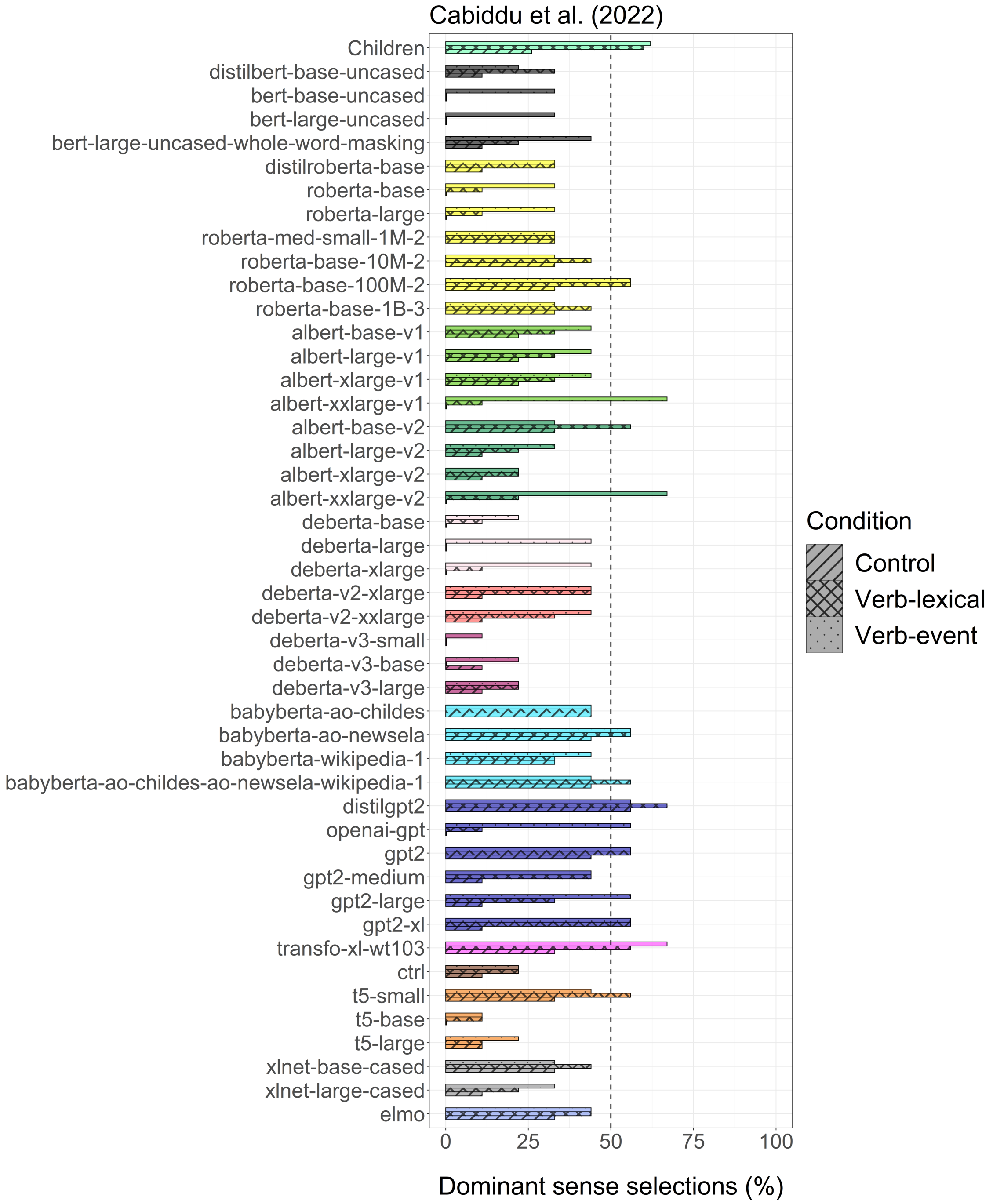

Figure S4.3. ***Percentage of dominant sense selections in models for Cabiddu et al. (2022), in dominant-plausible (Verb-lexical, Verb-event) and subordinate-plausible conditions (Control).***

Table S4.7. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the relative difference between models and children in Cabiddu et al. (2022), when considering performance in the Verb-Event structure condition. See our main paper for more details about this outcome measure. The predictors are log pretraining size, log model size, and their interaction. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	390.69	396.11	-192.34	384.69	-	-	-
+ Pretraining	4	390.37	397.60	-191.19	382.37	2.32	1	0.128
+ Model size	5	381.28	390.31	-185.64	371.28	11.09	1	**0.001**
+ Interaction	6	382.77	393.61	-185.39	370.77	0.51	1	0.477

Table S4.8. ***Output of the best model selected via model comparison in Table S4.7.***

		'+ Model size' model Verb-Event Condition Cabiddu et al. (2022)	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-50.56	-69.91 – -31.20	**<0.001**
Model size [log]	7.57	3.48 – 11.67	**0.001**
Pretraining size [log]	-0.30	-2.35 – 1.74	0.765
Random Effects			
σ^2	188.50		
$\tau_{00\ \text{family}}$	76.42		
ICC	0.29		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.300 / 0.502		

Table S4.9. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the relative difference between models and children in Cabiddu et al. (2022), when considering performance in the Verb-Lexical condition. See our main paper for more details about this outcome measure. The predictors are log pretraining size, log model size, and their interaction. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	340.21	345.63	-167.10	334.21	-	-	-
+ Pretraining	4	341.21	348.43	-166.60	333.21	1.00	1	0.317
+ Model size	5	341.23	350.27	-165.62	331.23	1.97	1	0.160
+ Interaction	6	342.00	352.84	-165.00	330.00	1.23	1	0.267

Table S4.10. ***Although no model surpassed the Null model in Table S4.9, below we show the output of the model including both main effects of model size and pretraining size, to appreciate size of the estimates and variance explained.***

	'+Model size' model Verb-Lexical Condition Cabiddu et al. (2022)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-30.39	-42.47 – -18.31	**<0.001**
Model size [log]	1.73	-0.87 – 4.34	0.186
Pretraining size [log]	0.16	-1.14 – 1.45	0.809
Random Effects			
σ^2	81.87		
$\tau_{00\ \mathrm{family}}$	22.74		
ICC	0.22		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.071 / 0.273		

References (for Appendix S4)

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/9kh29212

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, *49*, 1076–1089. https://doi.org/10.1037/a0026918

Appendix S5: Euclidean Distance Outcome Measure

In this section, we report results of the three experiments using an alternative outcome measure. See details about this measure in the main manuscript. In Figure S5.1, we show an example of how the measure is computed.

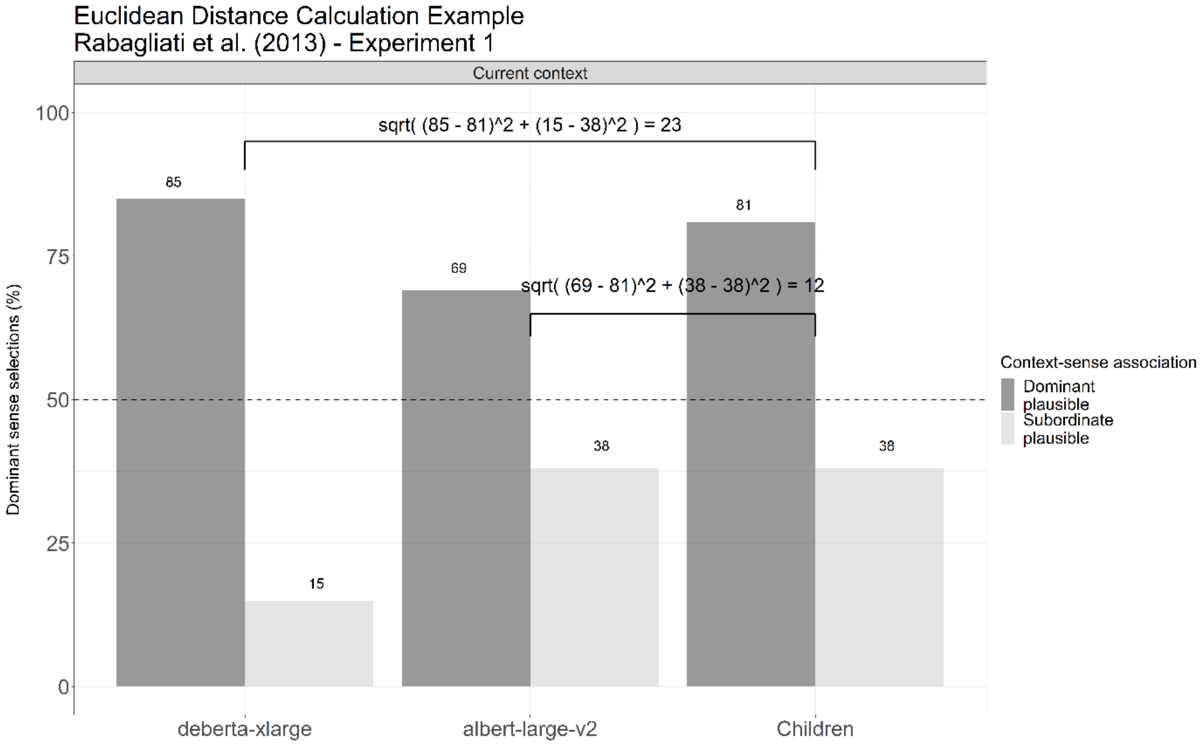

Figure S5.1. ***Example of calculation of the Euclidean Distance of deberta-xlarge and albert-large-v2 from children's scores in the Current Context condition of Rabagliati et al. (2013) experiment 1. The measure looks at the exact match between model and children.***

Rabagliati et al. (2013) – Experiment 1

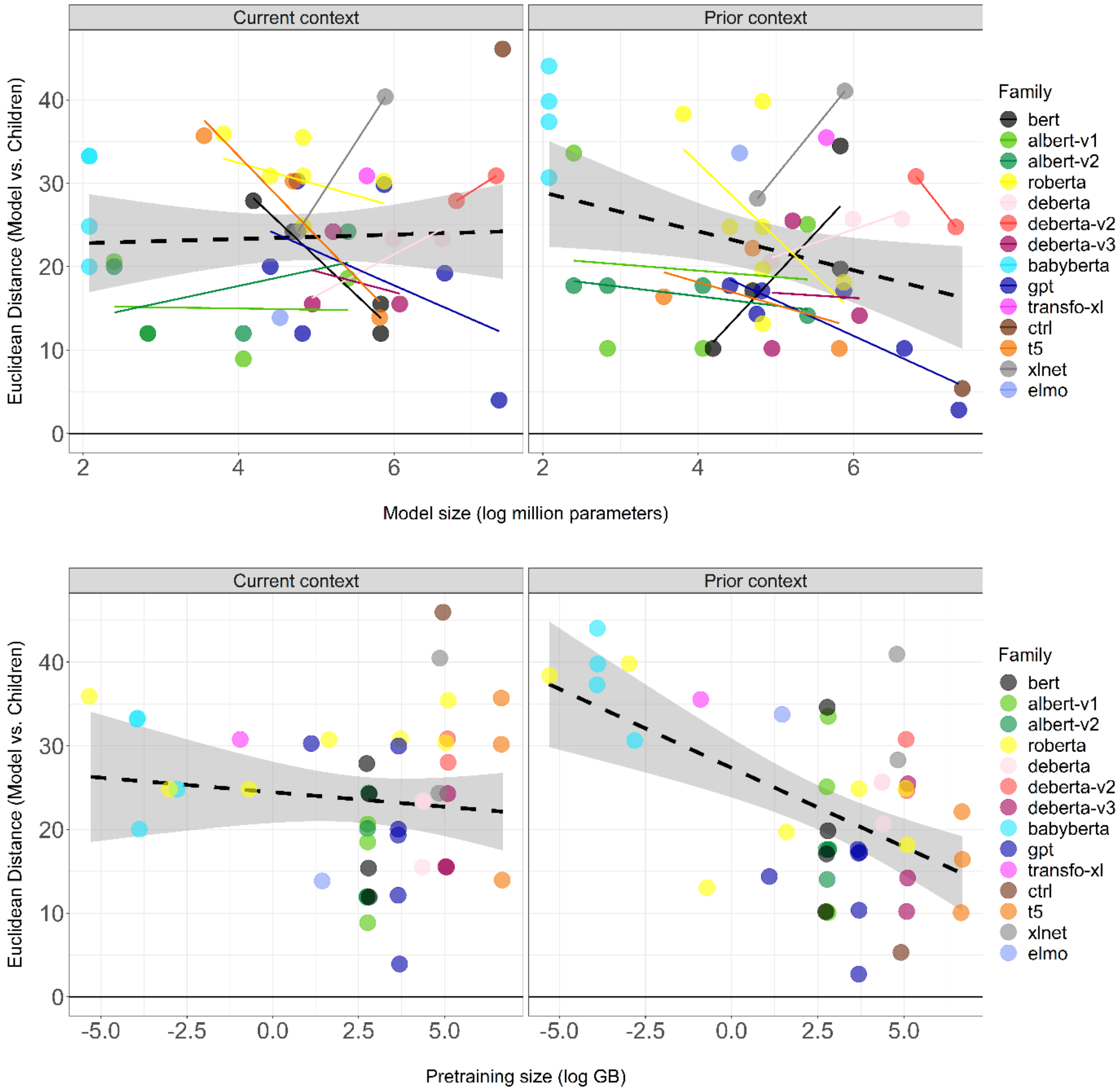

Figure S5.2. ***Models' Euclidean distance from children by model size (top row) and pretraining size (bottom row), in current and prior context conditions. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across models. Colored regression lines are also shown for each model family, although only when examining model size as there is almost null variation in pretraining size within family.***

Table S5.1. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Rabagliati et al. (2013) experiment 1. See our main paper for more details about this outcome measure. The predictors are dominant bias, condition (current, prior context), log pretraining size, log model size, and the pairwise interactions between model size, pretraining size, and condition. The random effect intercept is Model Family.***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	657.12	664.62	-325.56	651.12	-	-	-
+ Dominant Bias	4	658.23	668.23	-325.12	650.23	0.88	1	0.347
+ Condition	5	659.79	672.29	-324.89	649.79	0.44	1	0.505
+ Pretraining	6	646.98	661.98	-317.49	634.98	14.81	1	**0.000**
+ Model size	7	647.70	665.19	-316.85	633.70	1.28	1	0.257
+ Pretraining*Condition	8	641.30	661.30	-312.65	625.30	8.39	1	**0.004**
+ Size*Condition	9	642.39	664.89	-312.19	624.39	0.91	1	0.339
+ Pretraining*Model size	10	640.70	665.70	-310.35	620.70	3.68	1	0.055

Table S5.2. ***Output of the best model selected via model comparison in Table S5.1.***

	'Pretraining * Condition' model Euclidean Distance Rabagliati et al. (2013) - experiment 1		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	60.13	38.98 – 81.28	**<0.001**
Model size [log]	-0.95	-2.61 – 0.71	0.259
Pretraining size [log]	-1.03	-2.10 – 0.05	0.060
Condition [Prior context]	2.91	-1.31 – 7.12	0.173
Dominant bias	-0.57	-0.90 – -0.25	**0.001**
Pretraining size [log] * condition [Prior context]	-1.54	-2.60 – -0.48	**0.005**
Random Effects			
σ^2	56.96		
$\tau_{00\ \text{family}}$	13.78		
ICC	0.19		
N_{family}	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.271 / 0.413		

Rabagliati et al. (2013) – Experiment 2

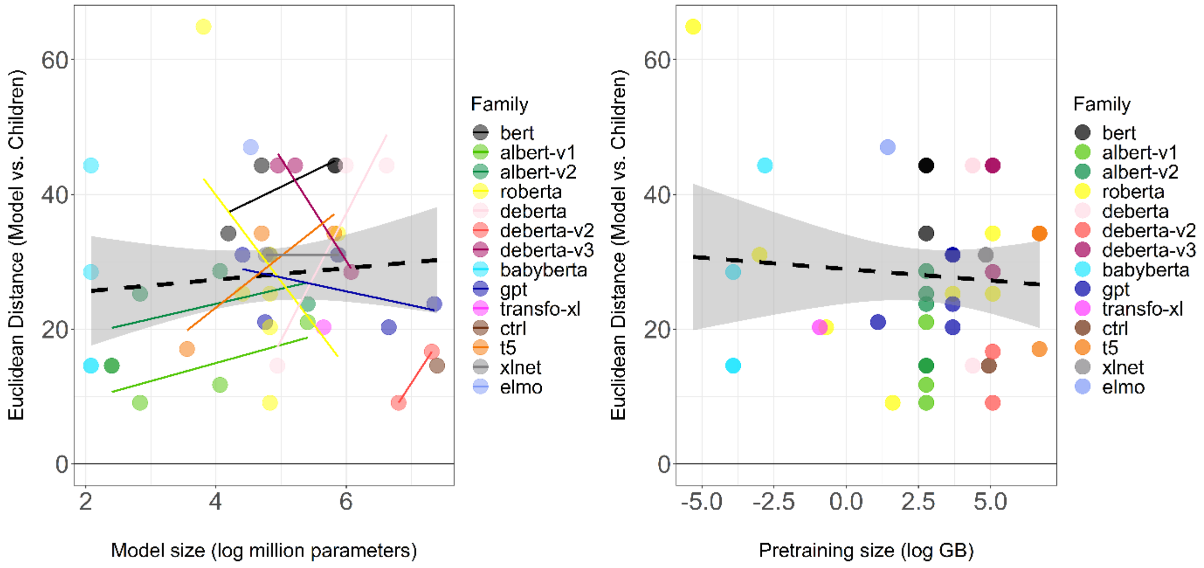

Figure S5.3. ***Models' Euclidean distance from children by model size and pretraining size in Rabagliati et al. (2013) experiment 2. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across models. Colored regression lines are also shown for each model family, although only when examining model size as there is almost null variation in pretraining size within family.***

Table S5.3. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Rabagliati et al. (2013) experiment 2. The predictors are dominant bias, log pretraining size, log model size, and the interaction between model size and pretraining size. The random effect intercept is Model Family.***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	359.62	365.04	-176.81	353.62	-	-	-
+ Dominant Bias	4	361.55	368.78	-176.78	353.55	0.07	1	0.789
+ Pretraining	5	362.38	371.42	-176.19	352.38	1.17	1	0.280
+ Model size	6	363.65	374.49	-175.82	351.65	0.73	1	0.391
+ Model size * Pretraining	7	365.63	378.28	-175.82	351.63	0.02	1	0.895

Table S5.4. ***Although no model surpassed the Null model in Table S5.3, below we show the output of the model including the main effects, to appreciate size of the estimates and variance explained.***

	'+ Model size' model Euclidean Distance Rabagliati et al. (2013) - Experiment 2		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	27.14	-16.95 – 71.23	0.221
Model size [log]	1.41	-2.08 – 4.90	0.418
Pretraining size [log]	-1.26	-3.23 – 0.70	0.202
Dominant bias	-0.05	-0.73 – 0.62	0.876
Random Effects			
σ^2	120.55		
$\tau_{00\ \text{family}}$	59.86		
ICC	0.33		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.049 / 0.365		

Cabiddu et al. (2022)

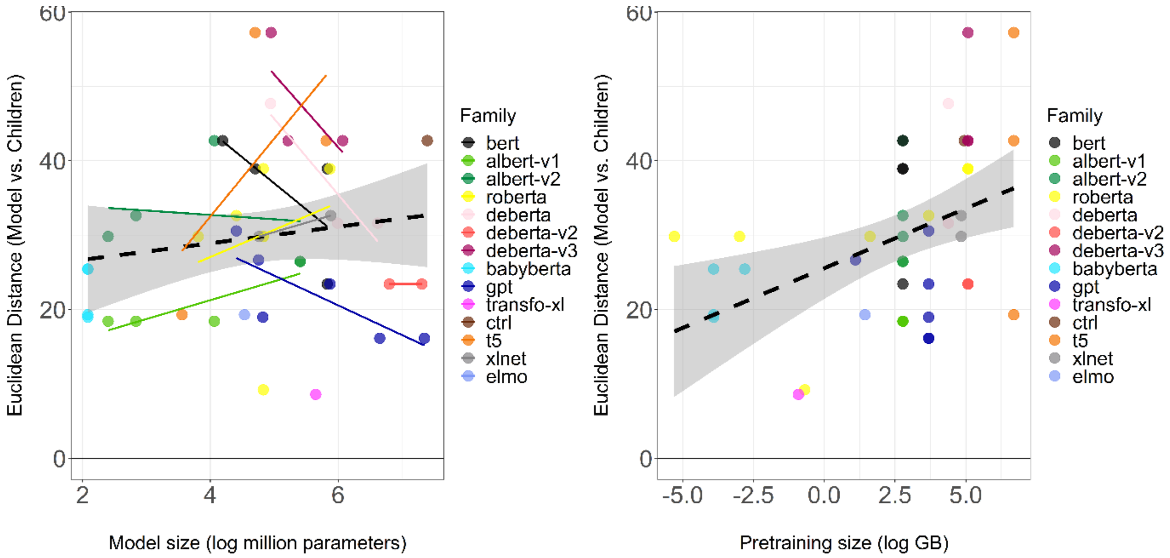

Figure S5.4. ***Models' Euclidean distance from children by model size and pretraining size, when comparing verb-event vs. control conditions in Cabiddu et al. (2022).***

Table S5.5. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Cabiddu et al. (2022) when comparing Verb-Event condition to Control. The predictors are dominant bias, log pretraining size, log model size, and the interaction between model size and pretraining size. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	344.69	350.11	-169.34	338.69	NA	NA	NA
+ Dominant Bias	4	343.94	351.17	-167.97	335.94	2.75	1	0.097
+ Pretraining	5	341.72	350.75	-165.86	331.72	4.22	1	**0.040**
+ Model size	6	343.04	353.88	-165.52	331.04	0.68	1	0.411
+ Model size * Pretraining	7	342.91	355.55	-164.45	328.91	2.14	1	0.144

Table S5.6. ***Output of the best model selected via model comparison in Table S5.5.***

	'+ Pretraining' model Euclidean Distance Verb-Event vs. Control Cabiddu et al. (2022)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	26.71	-1.86 – 55.28	0.066
Pretraining size [log]	1.54	0.01 – 3.07	**0.048**
Dominant bias	-0.02	-0.54 – 0.49	0.929
Random Effects			
σ^2	74.61		
$\tau_{00\ \text{family}}$	39.70		
ICC	0.35		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.167 / 0.456		

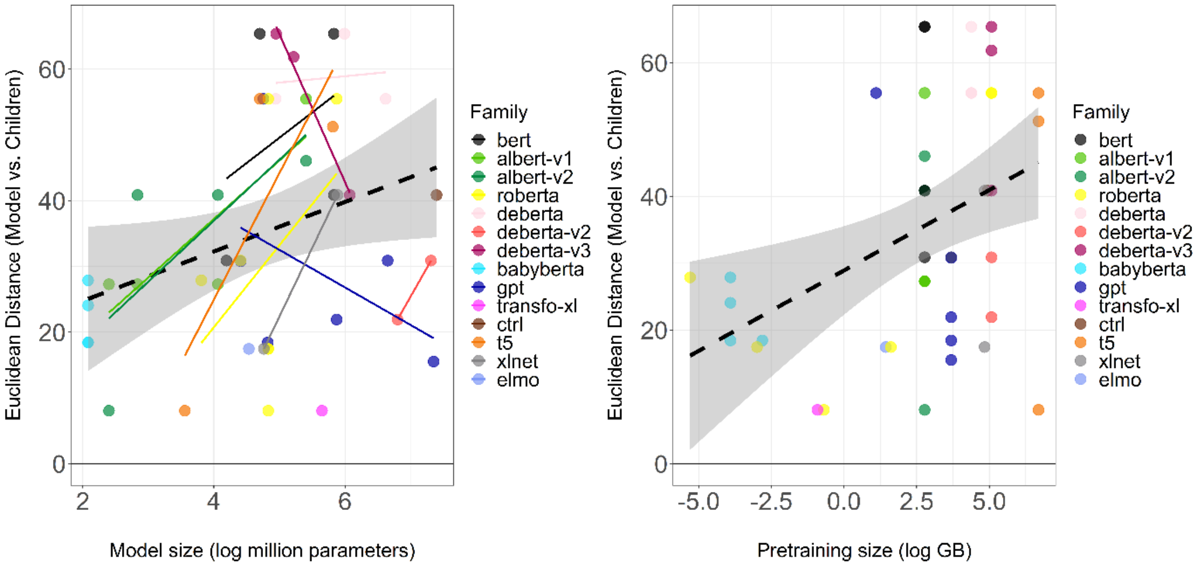

Figure S5.5. ***Models' Euclidean distance from children by model size and pretraining size, when comparing verb-lexical vs. control conditions in Cabiddu et al. (2022).***

Table S5.7. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Cabiddu et al. (2022) when comparing Verb-Lexical condition to Control. The predictors are dominant bias, log pretraining size, log model size, and the interaction between model size and pretraining size. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	390.83	396.25	-192.42	384.83	NA	NA	NA
+ Dominant Bias	4	389.72	396.95	-190.86	381.72	3.11	1	0.078
+ Pretraining	5	387.82	396.85	-188.91	377.82	3.90	1	**0.048**
+ Model size	6	389.03	399.87	-188.51	377.03	0.79	1	0.374
+ Model size * Pretraining	7	389.61	402.26	-187.81	375.61	1.41	1	0.234

Table S5.8. ***Output of the best model selected via model comparison in Table S5.7.***

	'+ Pretraining' model Euclidean Distance Verb-Lexical vs. Control Cabiddu et al. (2022b)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	32.85	-13.44 – 79.14	0.159
Pretraining size [log]	2.28	-0.13 – 4.69	0.063
Dominant bias	-0.08	-0.91 – 0.76	0.853
Random Effects			
σ^2	242.72		
$\tau_{00\ \text{family}}$	44.56		
ICC	0.16		
N_{family}	14		
Observations	45		
Marginal R^2 / Conditional R^2	0.158 / 0.289		

References (for Appendix S5)

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44). https://escholarship.org/uc/item/9kh29212

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology, 49*, 1076–1089. https://doi.org/10.1037/a0026918

Appendix S6: Simulations using adult-based sense prototypes

This section presents supplemental results obtained by using adult-directed speech to compute sense prototypes prior to testing the 45 Transformers in the word sense disambiguation tasks.

We initially explain how adult-directed speech was sense-tagged. Subsequently, we present plots showing the raw performance on the three experimental tasks for each adult-based Transformer.

This is followed by comparisons between adult-based models and the previously employed child-based models. For these comparisons, a preliminary examination was conducted to determine if adult-based models demonstrated superior performance than child-based models at the condition level, specifically looking at the percentage of correct responses given by a model in each experimental condition (note that this measure is independent of child performance).

Finally, we examined whether the child-based models better fit the children's data than the adult-based models. We begin by demonstrating that adult-based models did not display any dominance sense preference, thus highlighting the importance of using child-directed speech to derive child-based sense prototypes that reflect sense frequencies in the child input. We then show that child-based models better fit children's data in coherent tasks but not contrastive ones.

Sense Tagging the Spoken BNC

A question left open by previous analyses is whether the suboptimal performance of Transformers in contrastive tasks might be due to the use of sense prototypes computed from sense-tagged child-directed speech. Thus, it is possible that the models could perform better when their prototypes are based on adult-directed speech. Alternatively, the models may face difficulties with contrastive tasks for other reasons, such as a lack of real-world inference skills or multimodal data.

To build adult-based prototypes, we sense-tagged 80 utterances for each target word used in the study (40 utterances per sense). We extracted these utterances (available in our GitHub page) from adult-adult conversations present in the spoken section of the British National Corpus (BNC Consortium, 2007). One target word, *turkey*, had to be discarded because no utterances were available for one of its senses. For an additional four words, the input contained fewer than 40 utterances for one of the senses. Despite this, we used the number of utterances available and retained these target words in order to maximize the sample of items. In one case, a sense received a very low number of input utterances ($n = 3$). However, this was still retained on the basis that $n = 3$ is considered the minimum acceptable number to make sense prototypes functional in sense disambiguation (e.g., Loureiro et al., 2021). The frequencies of each tagged sense in the new adult input are displayed below in Table S6.1.

Table S6.1. ***For each target word's sense, the table displays the number of utterances tagged from the Spoken BNC.***

Target Word	Sense	n
band	music_group	40
band	object	40
bat	animal	9
bat	object	35
bow	knot	34
bow	weapon	3
button	clothing	40
button	tech	40
card	note	40
card	playing	40
chicken	animal	34
chicken	food	40
fish	animal	40
fish	food	40
glasses	drinking	40
glasses	eye	40
lamb	animal	24
lamb	food	40
letter	alphabet	40
letter	mail	40
line	geometry	40
line	order	40
nail	body_part	40
nail	object	40

Plots of Dominant Sense Selection – Raw Performance of Adult-Based Models

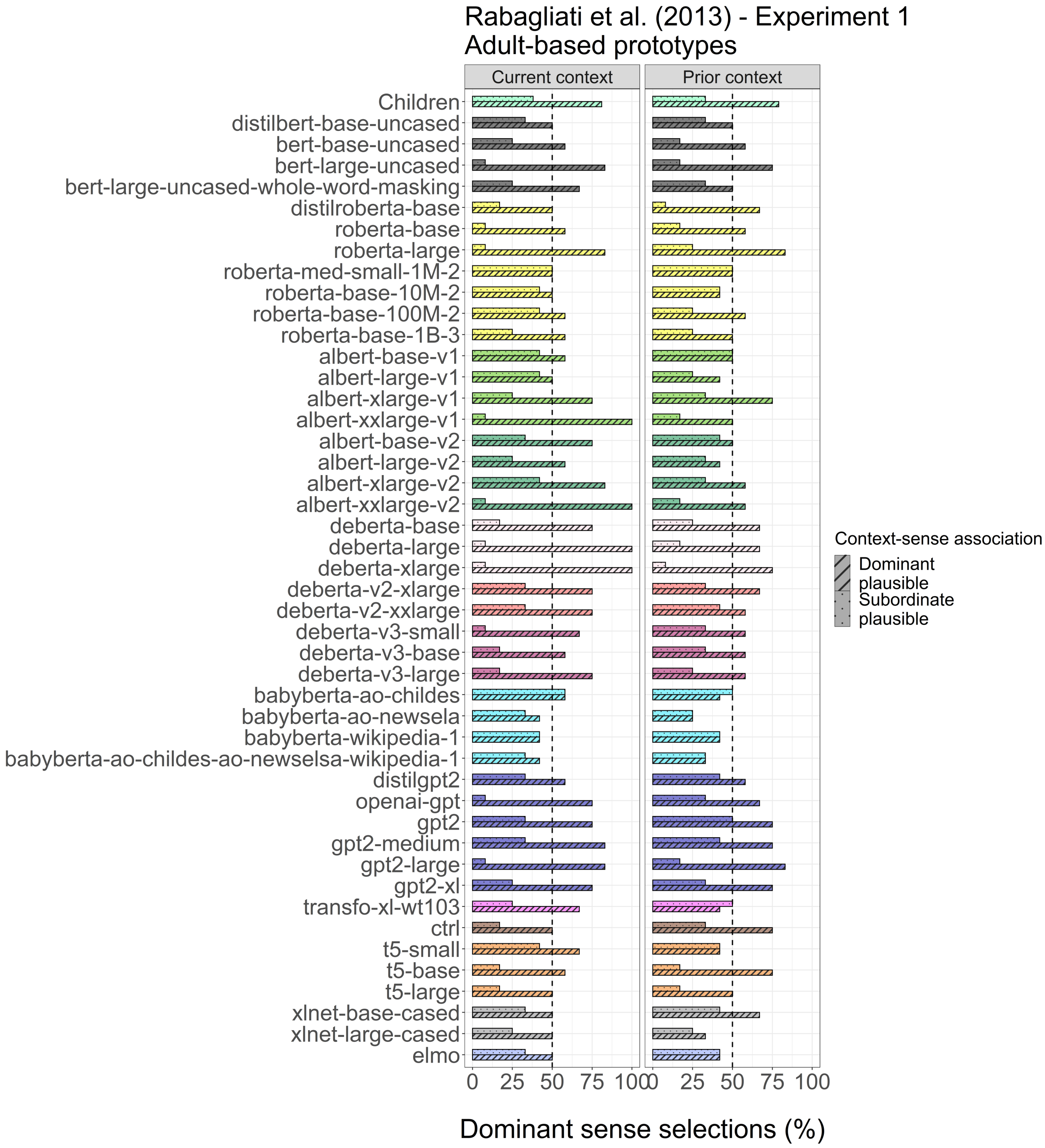

Figure S6.1. ***Percentage of dominant sense selections in adult-based models and children for experiment 1 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend), when disambiguation cues were included in current or prior context (left and right panel respectively).***

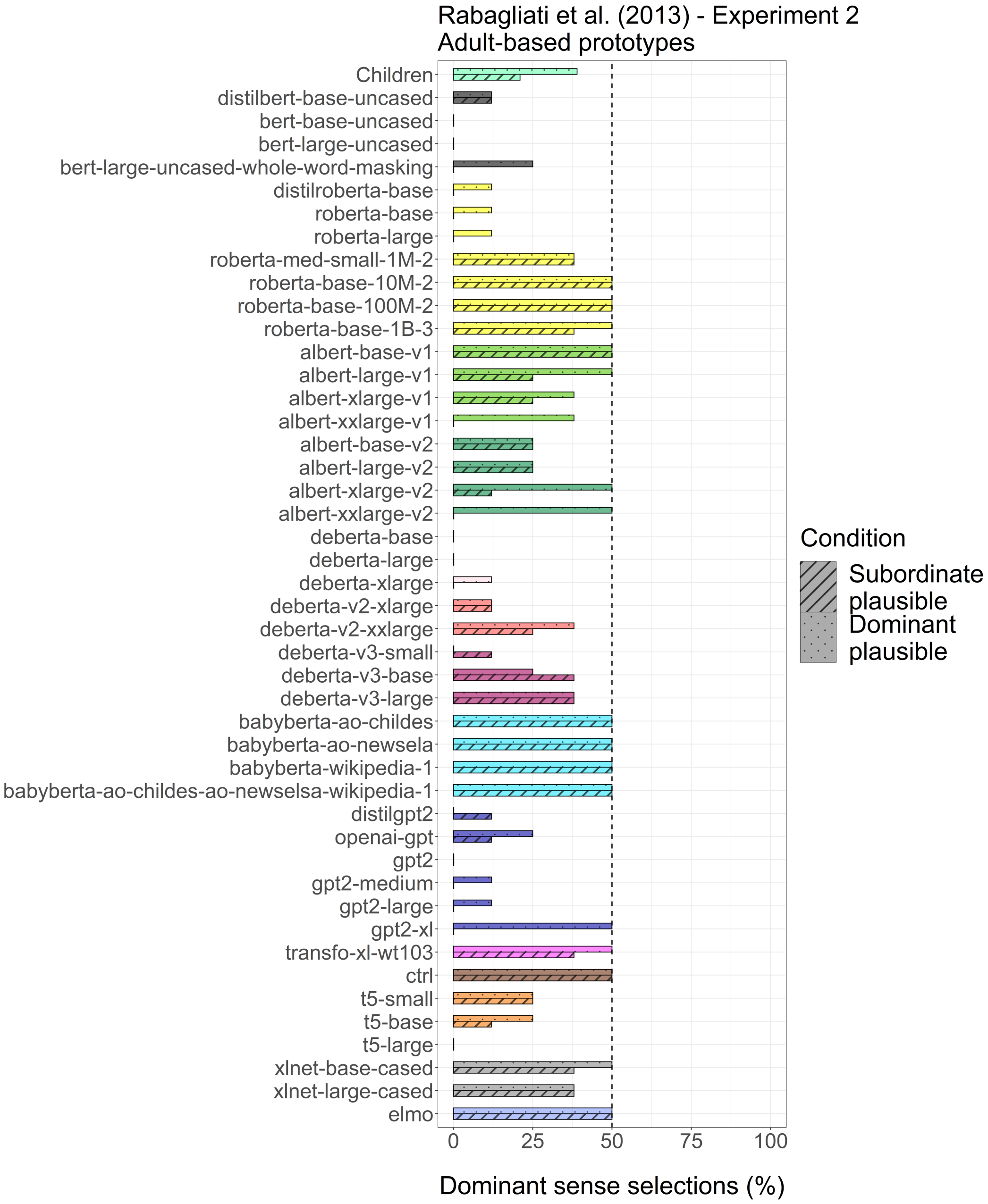

Figure S6.2. ***Percentage of dominant sense selections in adult-based models and children for experiment 2 of Rabagliati et al. (2013), in dominant-plausible and subordinate-plausible conditions (legend).***

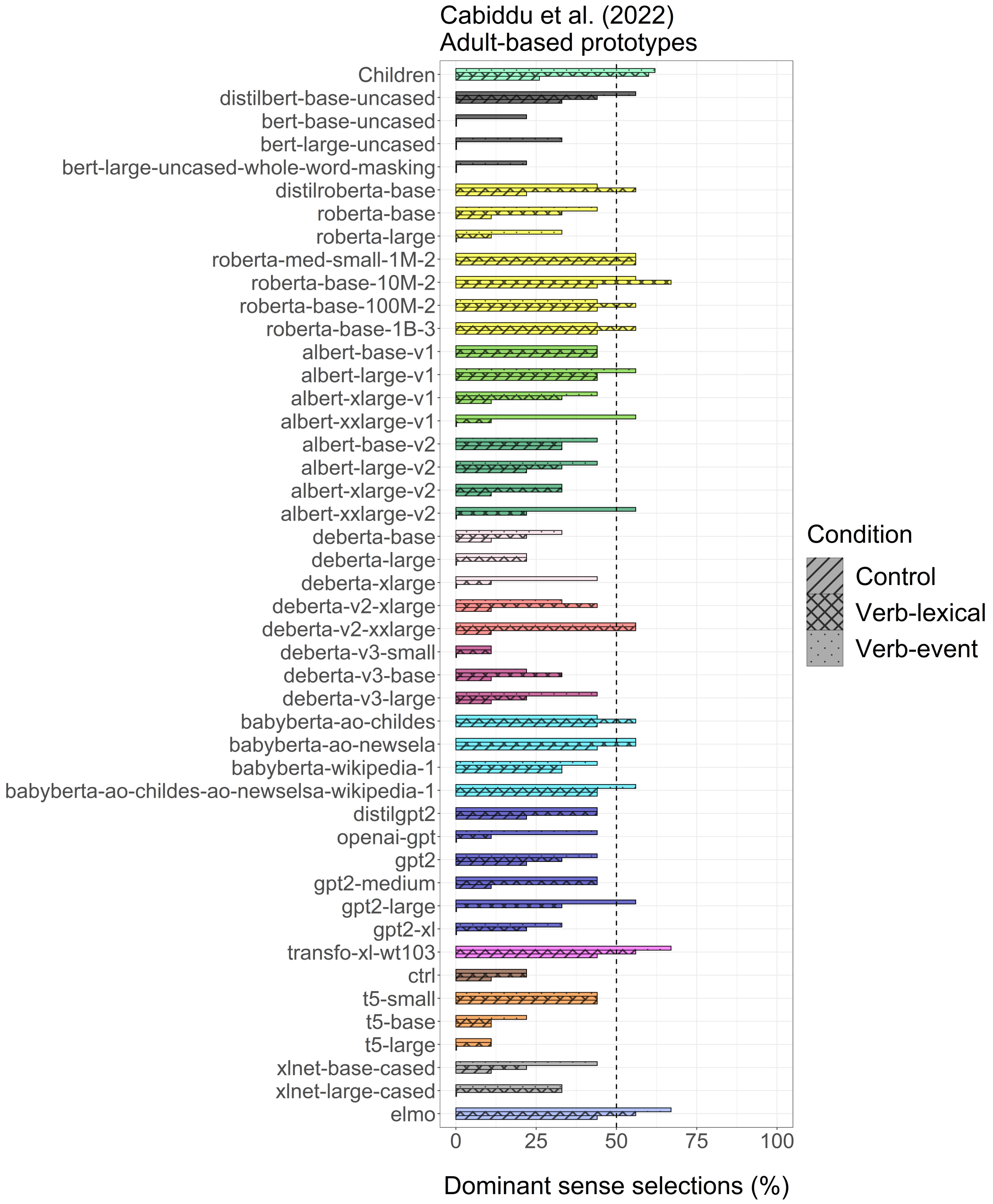

Figure S6.3. ***Percentage of dominant sense selections in adult-based models for Cabiddu et al. (2022), in dominant-plausible (Verb-lexical, Verb-event) and subordinate-plausible conditions (Control).***

Examining Correct Responses in Adult-Based and Child-Based Models

After implementing prototypes based on adult speech and rerunning the Transformers on the test stimuli, we examined the performance of the adult-based models in comparison to those child-based. As can be observed in Figure S6.4, the percentage of correct responses is remarkably similar across both age groups (adult-based models / child-based models) in each experiment considered. This reaffirms that the lower performance of Transformers in contrastive tasks, as seen in the Rabagliati Experiment 2 and Cabiddu Experiment 1, was not a consequence of deriving sense prototypes from child-directed speech.

It is important to note that this preliminary comparison does not take into account how closely the adult-based and child-based models approximate child performance. We relate the models' performance to child responses in the following section.

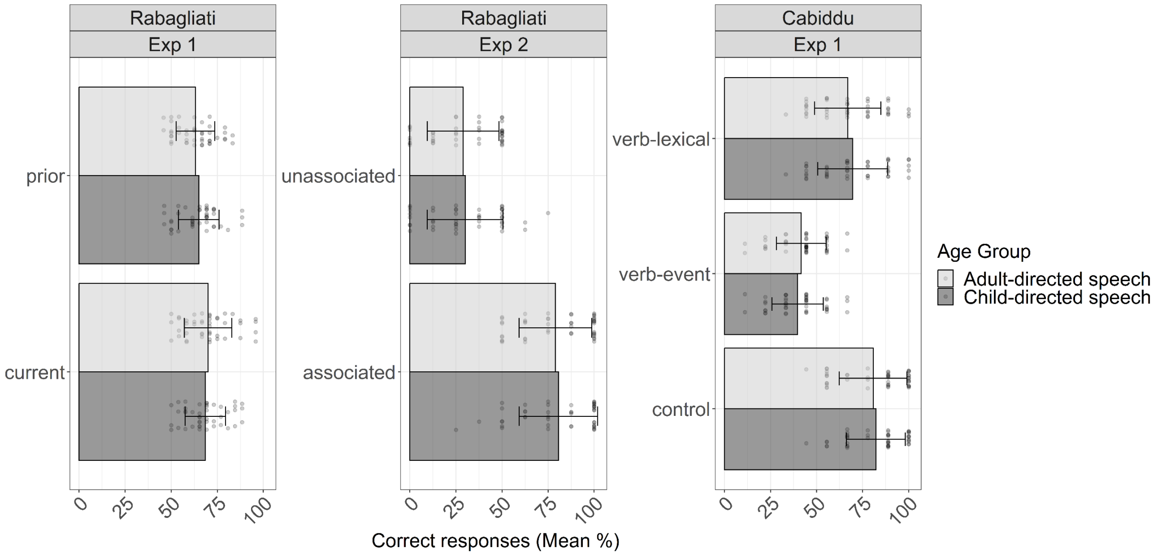

Figure S6.4. ***Mean percentage of correct responses (x-axis) in adult-based and child-based models (legend) for every condition (y-axis) in the behavioral experiments (panels). Error bars represent standard deviations around the mean percentages. Data points indicate performance for individual models in each condition.***

Relating adult-based and child-based models’ performance to child responses

Dominant Sense Preference

First, we investigated whether the models based on adult speech showed any preference for dominant senses (e.g., *elastic band*) or a subordinate sense (e.g., *music band*) in the initial experiment, which used coherent sentences (Rabagliati et al., 2013; Study 1).

In the experimental study involving both adults and children (Cabiddu et al., 2022), dominance had a more pronounced effect on child performance compared to adult performance. One hypothesis suggested that the distribution of sense frequencies might not be identical in adult-directed speech as it is in child-directed speech. If this hypothesis were correct, we would anticipate that Transformers would exhibit a weak or null dominance bias when their prototypes are derived from adult input. As shown in the figure S6.5, a visual comparison between adult-based and child-based dominance preference supports this expectation: The models did not display a dominance preference when using prototypes based on adult data. Further, we found a significant difference in dominance preference between adult-based and child-based models, in interaction with both model size and pretraining size (see Table S6.2 and S6.3). This finding supports the hypothesis that the dominant bias identified in the models based on child data was likely a result of employing sense-tagged child-directed speech.

Table S6.2. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the percentage of dominant senses selected across conditions of Rabagliati et al. (2013) experiment 1. The predictors are age group (adult-based model/child-based model), log pretraining size, log model size, and their interactions. Model family was used as random effect intercept. The Null model only includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	629.94	637.44	-311.97	623.94	NA	NA	NA
+ Age group	4	621.44	631.44	-306.72	613.44	10.50	1	**0.001**
+ Pretraining	5	613.08	625.58	-301.54	603.08	10.36	1	**0.001**
+ Model size	6	614.54	629.54	-301.27	602.54	0.54	1	0.461
+ Age group x Model size	7	590.41	607.91	-288.21	576.41	26.13	1	**0.000**
+ Age group x Pretraining	8	587.07	607.07	-285.53	571.07	5.34	1	**0.021**
+ Pretraining x Model size	9	589.05	611.55	-285.52	571.05	0.02	1	0.884
+ Age group x Pretraining x Model size	10	587.48	612.47	-283.74	567.48	3.57	1	0.059

Table S6.3. ***Output of the best model selected via model comparison in Table S6.2.***

	Dominant Sense Preference **'+ Age group * Pretraining' model**		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	39.67	32.69 – 46.65	**<0.001**
Age group [Child-directed speech]	23.82	15.26 – 32.38	**<0.001**
Model size [log]	1.29	-0.23 – 2.80	0.096
Pretraining size [log]	-0.35	-1.10 – 0.40	0.354
Age group [Child-directed speech] × Model size [log]	-3.34	-5.28 – -1.39	**0.001**
Age group [Child-directed speech] × Pretraining size [log]	-1.08	-2.03 – -0.14	**0.025**
Random Effects			
σ^2	31.66		
$\tau_{00\ \text{family}}$	6.36		
ICC	0.17		
N_{family}	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.428 / 0.524		

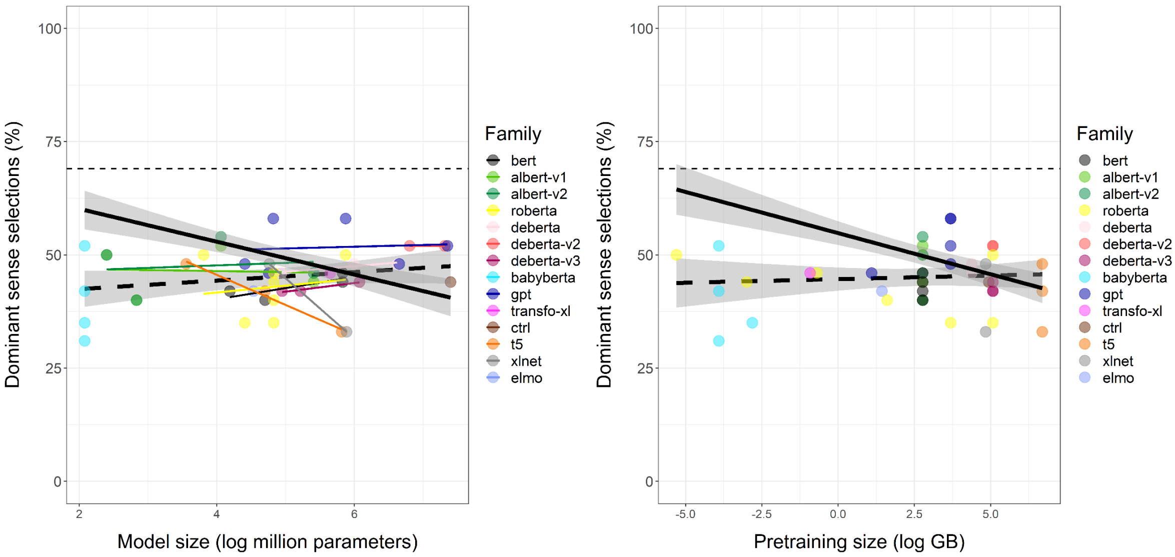

Figure S6.5. ***The percentage of dominant sense selections by adult-based models, in Rabagliati et al. (2013; Study 1), is randomly distributed around 50% and never reaches the level of child dominance bias (indicated by the dashed horizontal line). Furthermore, the selections of dominant senses do not change as a function of either the model size or the pretraining size. The solid lines display the dominant sense selection patterns in the child-based models for a visual comparison.***

Euclidean Distance Measure

In this section, we examined whether child-based models fit children's responses better than adult-based models in each of the three experiments. We used the measure of Euclidean Distance that, as presented in Appendix S5, evaluates the exact match between the model and the child.

To foresee, the only significant difference between adult-based and child-based models was found when examining performance in resolving coherent stories (Rabagliati et al., 2013; Study 1).

In Figure S6.6, we show that child-based models performed more closely to child performance than the adult models did in the first experiment. This can be observed by examining the differences between adult-based dashed regression lines and child-based solid regression lines, with child-based models' regression lines being closer to child performance ($y = 0$). The difference in Euclidean distance from children between adult-based models and child-based models was significant, as shown in table S6.4 and S6.5.

For what concerns the contrastive tasks, instead, tables S6.6 to S6.11 show non-significant differences between adult-based models and child-based models at capturing child performance.

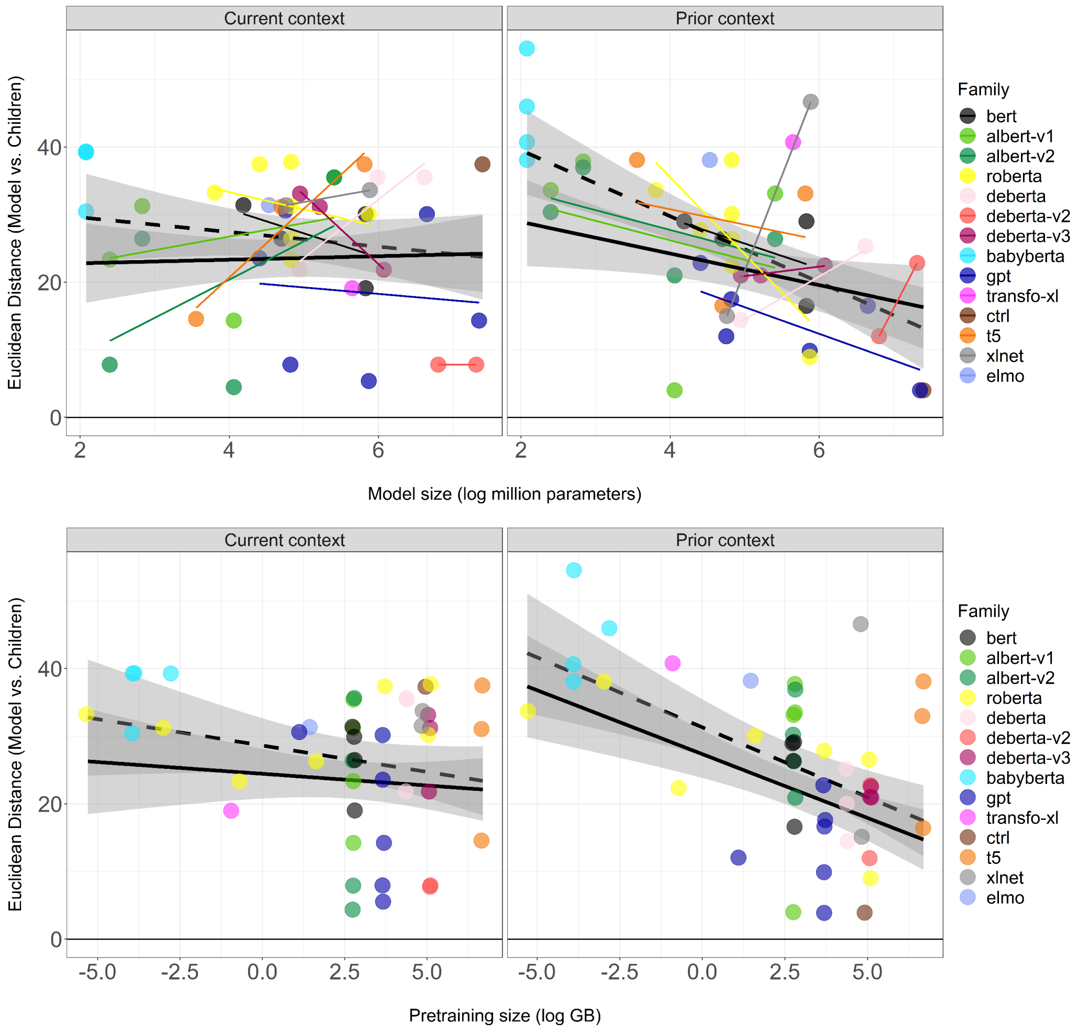

Figure S6.6. ***Models' Euclidean distance from children by model size (top row) and pretraining size (bottom row), in current and prior context conditions of Rabagliati et al. (2013), experiment 1. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across adult-based models. The solid regression line with 95% confidence interval shows performance across child-based models. Colored regression lines are also shown for each adult-based model family, although only when examining model size as there is almost null variation in pretraining size within family. Colored data points refer to the adult-based dataset. Data points from child-based dataset are omitted for simplicity.***

Table S6.4. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Rabagliati et al. (2013) experiment 1. See our main paper for more details about this outcome measure. The predictors are age group (adult-based model/child-based model), condition (current, prior context), log pretraining size, log model size, and their two-way and three-way interactions. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	1333.28	1342.86	-663.64	1327.28	NA	NA	NA
+ Age group	4	1329.41	1342.19	-660.71	1321.41	5.87	1	**0.015**
+ Condition	5	1330.94	1346.90	-660.47	1320.94	0.47	1	0.491
+ Pretraining	6	1324.41	1343.56	-656.20	1312.41	8.53	1	**0.003**
+ Model size	7	1325.45	1347.80	-655.73	1311.45	0.95	1	0.329
+ Age group x Condition	8	1327.42	1352.97	-655.71	1311.42	0.03	1	0.862
+ Age group x Model size	9	1324.78	1353.52	-653.39	1306.78	4.64	1	**0.031**
+ Age group x Pretraining	10	1326.43	1358.36	-653.22	1306.43	0.35	1	0.556
+ Condition x Pretraining	11	1317.48	1352.60	-647.74	1295.48	10.95	1	**0.001**
+ Condition x Model size	12	1314.56	1352.88	-645.28	1290.56	4.92	1	**0.027**
+ Pretraining x Model size	13	1315.14	1356.65	-644.57	1289.14	1.42	1	0.233
+ Age group x Condition x Pretraining	14	1317.05	1361.75	-644.52	1289.05	0.09	1	0.759
+ Age group x Condition x Model size	15	1317.95	1365.85	-643.98	1287.95	1.09	1	0.296
+ Age group x Pretraining x Model size	16	1319.94	1371.02	-643.97	1287.94	0.02	1	0.891

Table S6.5. ***Output of the best model selected via model comparison in Table S6.4.***

	‘Condition x Model size’ model Euclidean Distance Rabagliati et al. (2013) - experiment 1		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	31.32	21.92 – 40.71	**<0.001**
Age group [Child-directed speech]	-13.22	-22.68 – -3.77	**0.006**
Condition [Prior context]	12.38	2.93 – 21.84	**0.011**
Model size [log]	-0.67	-2.67 – 1.32	0.506
Pretraining size [log]	-0.30	-1.28 – 0.69	0.554
Age group [Child-directed speech] × Condition [Prior context]	-0.46	-5.48 – 4.57	0.858
Age group [Child-directed speech] × Pretraining size [log]	-0.31	-1.31 – 0.70	0.548
Age group [Child-directed speech] × Model size [log]	2.29	0.22 – 4.37	**0.030**
Condition [Prior context] × Pretraining size [log]	-0.80	-1.81 – 0.20	0.116
Condition [Prior context] × Model size [log]	-2.29	-4.36 – -0.21	**0.031**
Random Effects			
σ^2	72.92		
$\tau_{00\ \text{family}}$	16.61		
ICC	0.19		
N_{family}	14		
Observations	180		
Marginal R^2 / Conditional R^2	0.195 / 0.345		

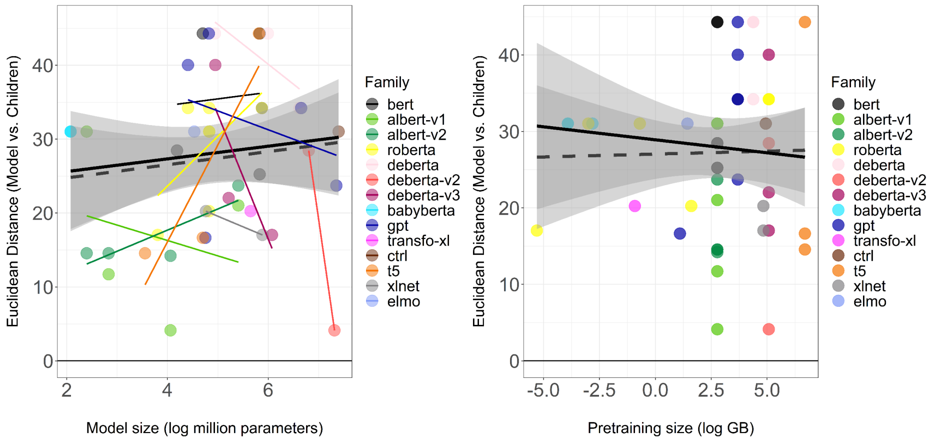

Figure S6.7. ***Models' Euclidean distance from children by model size and pretraining size in Rabagliati et al. (2013) experiment 2. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across adult-based models. The solid regression line with 95% confidence interval shows performance across child-based models. Colored regression lines are also shown for each adult-based model family, although only when examining model size as there is almost null variation in pretraining size within family. Colored data points refer to the adult-based dataset. Data points from child-based dataset are omitted for simplicity.***

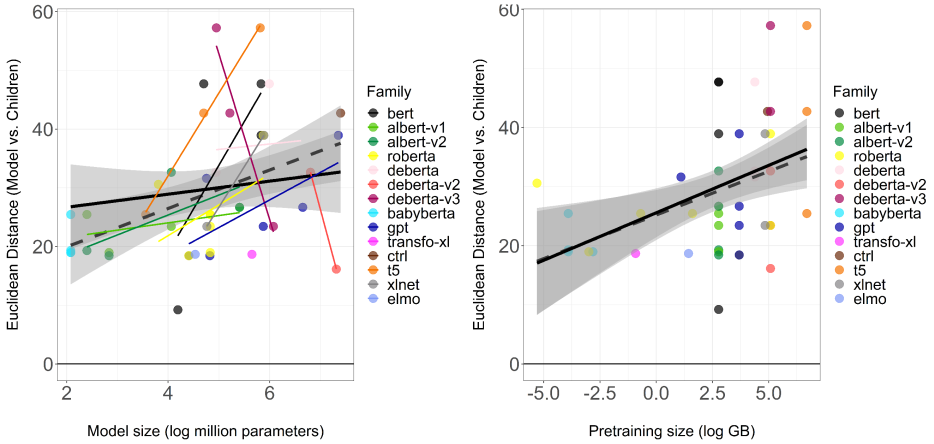

Figure S6.8. ***Models' Euclidean distance from children by model size and pretraining size in Cabiddu et al. (2022), Verb-Event condition. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across adult-based models. The solid regression line with 95% confidence interval shows performance across child-based models. Colored regression lines are also shown for each adult-based model family, although only when examining model size as there is almost null variation in pretraining size within family. Colored data points refer to the adult-based dataset. Data points from child-based dataset are omitted for simplicity.***

Table S6.6. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Rabagliati et al. (2013) experiment 2. See our main paper for more details about this outcome measure. The predictors are age group (adult-based model/child-based model), log pretraining size, log model size, and interactions. The random effect intercept is Model Family. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Df**	**Pr(>Chisq)**
Null model	3	693.46	700.96	-343.73	687.46	NA	NA	NA
+ Age group	4	695.33	705.33	-343.66	687.33	0.13	1	0.716
+ Pretraining	5	697.21	709.71	-343.60	687.21	0.12	1	0.728
+ Model size	6	698.79	713.79	-343.40	686.79	0.41	1	0.520
+ Age group x Model size	7	700.79	718.29	-343.40	686.79	0.00	1	0.983
+ Age group x Pretraining	8	702.32	722.31	-343.16	686.32	0.48	1	0.490
+ Pretraining x Model size	9	703.57	726.07	-342.78	685.57	0.75	1	0.387
+ Age group x Pretraining x Model size	10	704.57	729.57	-342.29	684.57	0.99	1	0.319

Table S6.7. ***Although no model surpassed the Null model in Table S6.6, below we show the output of the model including the main effects, to appreciate size of the estimates and variance explained.***

	'+ Model size' model - Euclidean Distance Rabagliati et al. (2013) - Experiment 2		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	24.31	12.79 – 35.83	**<0.001**
Age group [Child-directed speech]	0.77	-3.51 – 5.05	0.721
Model size [log]	0.70	-1.58 – 2.99	0.542
Pretraining size [log]	-0.32	-1.47 – 0.83	0.581
Random Effects			
σ^2	104.19		
$\tau_{00\ family}$	44.62		
ICC	0.30		
N $_{family}$	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.007 / 0.305		

Table S6.8. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Cabiddu et al. (2022) Verb-Event condition. See our main paper for more details about this outcome measure. The predictors are age group (adult-based model/child-based model), log pretraining size, log model size, and interactions. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	678.16	685.66	-336.08	672.16	NA	NA	NA
+ Age group	4	680.02	690.02	-336.01	672.02	0.14	1	0.706
+ Pretraining	5	675.80	688.30	-332.90	665.80	6.22	1	**0.013**
+ Model size	6	676.18	691.18	-332.09	664.18	1.62	1	0.203
+ Age group x Model size	7	675.29	692.79	-330.64	661.29	2.89	1	0.089
+ Age group x Pretraining	8	675.34	695.34	-329.67	659.34	1.94	1	0.163
+ Pretraining x Model size	9	673.64	696.14	-327.82	655.64	3.71	1	0.054
+ Age group x Pretraining x Model size	10	675.63	700.63	-327.81	655.63	0.01	1	0.918

Table S6.9. ***Output of the best model selected via model comparison in Table S6.8.***

	'+ Pretraining' model -Euclidean Distance Verb-Event Condition -Cabiddu et al. (2022)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	25.80	20.83 – 30.77	**<0.001**
Age group [Child-directed speech]	0.71	-3.08 – 4.51	0.710
Pretraining size [log]	1.25	0.32 – 2.18	**0.009**
Random Effects			
σ^2	82.01		
$\tau_{00\ \text{family}}$	33.36		
ICC	0.29		
N $_{\text{family}}$	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.110 / 0.367		

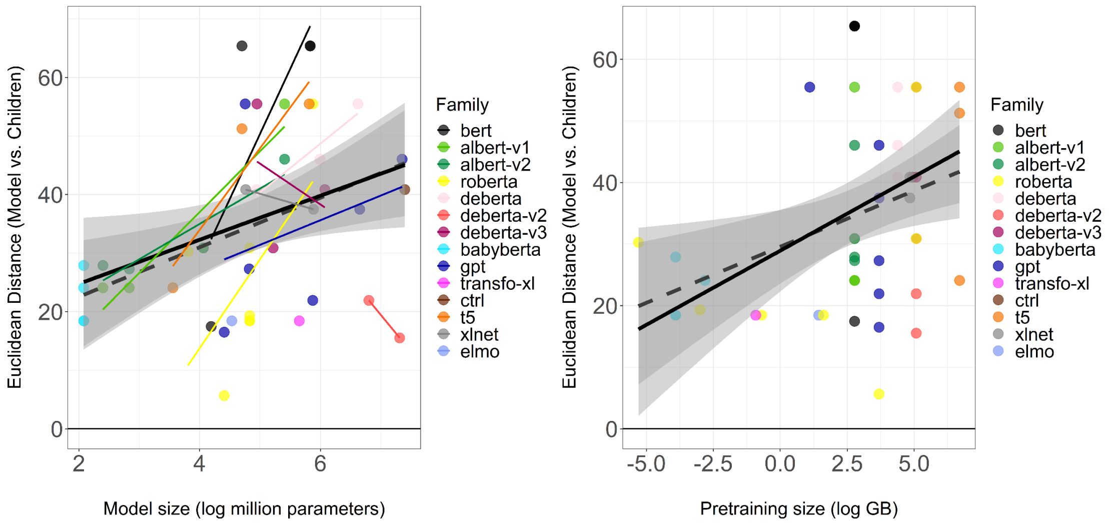

Figure S6.9. ***Models' Euclidean distance from children by model size and pretraining size in Cabiddu et al. (2022), Verb-Lexical condition. Model families are shown in the legend. The black horizontal line (y=0) indicates child performance. The dashed regression line with 95% confidence interval shows performance across adult-based models. The solid regression line with 95% confidence interval shows performance across child-based models. Colored regression lines are also shown for each adult-based model family, although only when examining model size as there is almost null variation in pretraining size within family. Colored data points refer to the adult-based dataset. Data points from child-based dataset are omitted for simplicity.***

Table S6.10. ***Model comparison between nested linear mixed-effect models via likelihood ratio test. The outcome is the Euclidean Distance between models and children in Cabiddu et al. (2022) Verb-Lexical condition. See our main paper for more details about this outcome measure. The predictors are age group (adult-based model/child-based model), log pretraining size, log model size, and interactions. The random effect intercept is Model Family. The Null model includes main and random effect intercepts. Subsequent models add one predictor at a time. The table shows the number of model parameters (npar), Akaike (AIC) and Bayesian (BIC) Information criterions, log-likelihood (logLik), deviance, Chi-square statistic (Chisq), degrees of freedom (Df), and p value Pr(>Chisq).***

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Null model	3	755.38	762.88	-374.69	749.38	NA	NA	NA
+ Age group	4	757.29	767.29	-374.65	749.29	0.09	1	0.760
+ Pretraining	5	752.51	765.01	-371.26	742.51	6.78	1	**0.009**
+ Model size	6	746.59	761.59	-367.29	734.59	7.92	1	**0.005**
+ Age group x Model size	7	748.52	766.02	-367.26	734.52	0.06	1	0.800
+ Age group x Pretraining	8	749.62	769.62	-366.81	733.62	0.90	1	0.343
+ Pretraining x Model size	9	748.72	771.22	-365.36	730.72	2.90	1	0.088
+ Age group x Pretraining x Model size	10	750.33	775.33	-365.17	730.33	0.39	1	0.534

Table S6.11. ***Output of the best model selected via model comparison in Table S6.10.***

	'+ Model size' model -Euclidean Distance Verb-Lexical Condition Cabiddu et al. (2022)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	7.66	-8.10 – 23.42	0.336
Age group [Child-directed speech]	0.91	-4.54 – 6.36	0.742
Model size [log]	4.80	1.76 – 7.83	**0.002**
Pretraining size [log]	0.86	-0.68 – 2.39	0.270
Random Effects			
σ^2	168.96		
$\tau_{00\ \text{family}}$	110.41		
ICC	0.40		
N_{family}	14		
Observations	90		
Marginal R^2 / Conditional R^2	0.214 / 0.524		

References (for Appendix S6)

BNC Consortium. (2007). *British National Corpus, XML edition.* https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022). The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/9kh29212

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). *Analysis and Evaluation of Language Models for Word Sense Disambiguation* (arXiv:2008.11608). arXiv. https://doi.org/10.48550/arXiv.2008.11608

Rabagliati, H., Pylkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology, 49,* 1076–1089. https://doi.org/10.1037/a0026918

License

Language Development Research (ISSN 2771-7976) is published by TalkBank and the Carnegie Mellon University Library Publishing Service. Copyright © 2024 The Author(s). This work is distributed under the terms of the Creative Commons Attribution-Noncommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/), which permits any use, reproduction and distribution of the work for noncommercial purposes without further permission provided the original work is attributed as specified under the terms available via the above link to the Creative Commons website.