Comparing Children and Large Language Models in Word Sense Disambiguation: Insights and Challenges

Francesco Cabiddu[1] (cabidduf@cardiff.ac.uk)
Mitja Nikolaus[2,3] (mitja.nikolaus@univ-amu.fr)
Abdellah Fourtassi[2] (abdellah.fourtassi@univ-amu.fr)

[1]Cardiff University, School of Psychology, Cardiff, United Kingdom
[2]Aix Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
[3]Aix-Marseille Université, CNRS, LPL, Aix-en-Provence, France

Abstract

Understanding how children process ambiguous words is a challenge because sense disambiguation depends on sentence context bottom-up and top-down aspects. 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. 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 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

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 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 (e.g., Cabiddu, Bott, Jones, & Gambi, 2022b; Rabagliati, Pylkkänen, & Marcus, 2013). That is, how children use sense-specific representations (e.g., band = music band, elastic band).

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 syntactic and semantic sentence structures (e.g., Jawahar, Sagot, & Seddah, 2019; Tenney, Das, & Pavlick, 2019) on which sense disambiguation depends. Here, we refer to these high-level structures as top-down cues that a usage-based learner might acquire through language experience (Alishahi & Stevenson, 2013; Bybee, 2010). Transformers' inherent sensitivity to these cues allowed us to apply these models to raw naturalistic language, without having to enrich the input with external resources to provide sensitivity to such structures (e.g., Alishahi & Stevenson, 2013).

Transformers form adult-like sense representations in natural language classification tasks, where annotators select a fitting target sense given the sentence context (Loureiro, Rezaee, Pilehvar, & Camacho-Collados, 2021). However, these 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 adult-like from child-like performance, as both adults and children perform well at disambiguating coherent sentences (e.g., Khanna & Boland, 2010; Rabagliati et al., 2013). Thus, we tested 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 have computed models' representations based on adult language (Loureiro, Jorge, & Camacho-Collados, 2022; Loureiro et al., 2021). Here, we evaluated how properties of child sense processing could be captured using model representations formed from naturalistic child-directed speech. This choice is motivated by differences in how senses are assigned to words in children and adults, likely due to differences in word use in child and adult environments (Meylan, Mankewitz, Floyd, Rabagliati, & Srinivasan, 2021).

We examined Transformers using datasets from 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 (Schrimpf et al., 2021) allowed us to examine consistent dimensions across models, instead of relying on single models which might be influenced by specific peculiarities (architecture, pretraining objectives, amount/type of pretraining input, etc.). Specifically, we ex-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

plored how scalability in models' size (number of parameters) and pretraining data size related to sense disambiguation. These dimensions have been explored separately and not specifically in child sense disambiguation. Based on findings about word age of acquisition norms (Laverghetta Jr & Licato, 2021), we expected models with a higher number of parameters to better fit child data. We also expected a null effect of pretraining size, in line with 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).

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 master word ambiguities in naturalistic conversations (Meylan et al., 2021), which raises a question about which sentence properties facilitate child word disambiguation (Cabiddu et al., 2022b; Hahn, Snedeker, & Rabagliati, 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 considered. 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, we investigated 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.

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 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). However, a significant difference from a control condition emerged (21% dominant selections when the story

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%

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.

Cabiddu et al. (2022b) focused on verbs. 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 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, Daum, & Huettig, 2016), this condition tested if children would weigh more word associations coming from a verb than the rest of the context.

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

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, Kidd, Rowland, & Theakston, 2015; McCauley & Christiansen, 2019; Saffran, Aslin, & Newport, 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 idea of context-dependent representations aligns 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., 2022, 2021). In Loureiro et al. (2021), Transformers were evaluated using a nearest neighbor approach (e.g., Melamud, Goldberger, & Dagan, 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 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, Remus, Chawla, & Biemann, 2019). Model descriptions can be found on our OSF page, where we also share materials and code to reproduce the study results (`https://doi.org/10.17605/OSF.IO/A2BZQ`).

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). On our OSF, 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*). 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*.

To evaluate model performance, we extracted a contextualized vector for each test sentence's target word. We compared each vector to the two sense prototypes using cosine similarity. 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, Bott, Jones, & Gambi, 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*). 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. Details about items and annotation process are on our OSF. 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. 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.

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.

For example, in the first experiment, children selected dominant senses (e.g., *elastic band*) in 81% of trials in the dominant-plausible condition (*She streched 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.

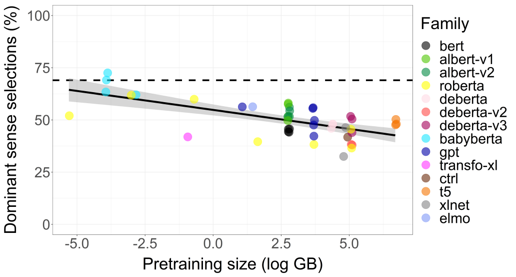

Figure 1: Percentage of dominant senses selected by each model in Rabagliati et al. (2013) experiment 1, by pretraining size in log GB. The dashed horizontal line indicates dominant sense prevalence in ChiSense-12.

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 nested linear mixed-effects model indicated higher context sensitivity as model size increased (β = 5.36, *95% CI* = [2.07, 8.64], *p* = .002) and pretraining size increased (β = 3.81, *95% CI* = [2.16, 5.47], *p* <.001). A main effect of condition (β = -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).

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, 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.

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 (β = 7.57, *95% CI* = [3.48, 11.67], *p* = .001), but not pretraining size (β = -.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 (β = 1.73, *95% CI* = [-0.87, 4.34], *p* = .186) or pretraining size (β = 0.16, *95% CI* = [-1.14, 1.45], *p* = .809).

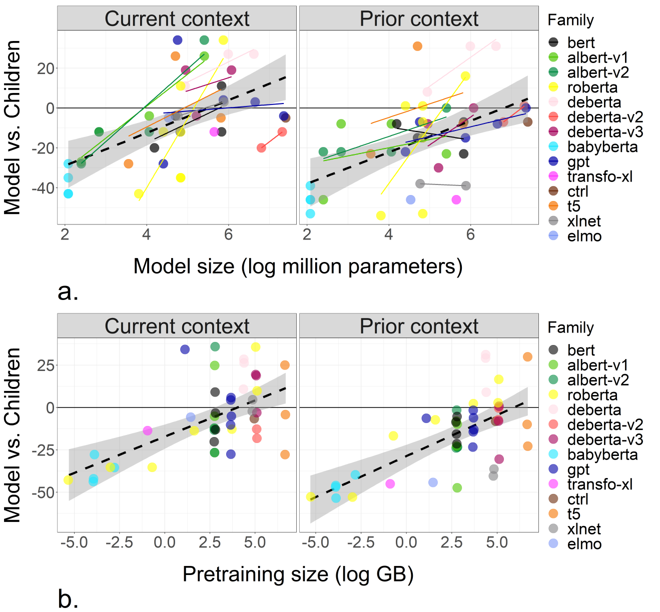

Figure 2: Models' distance from children by model size (a) and pretraining size (b), in current and prior context conditions. Model families are shown in the legend. The black horizontal line indicates child performance. The dashed regression line with 95% confidence interval shows performance across 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.

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 sense knowledge can, in principle, arise from probabilistic representations embedding context-dependent and context-independent information (Ambridge, 2020; Srinivasan & Rabagliati, 2021).

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, Chang, Lee, & Toutanova, 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.

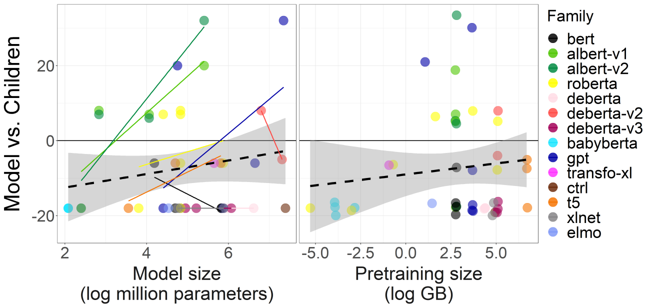

Figure 3: Models' distance from children by model size and pretraining size, in Rabagliati et al. (2013) experiment 2.

Further, a visual inspection of models' performance at contrastive tasks (see raw plots of sense selection for each model in OSF, 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 OSF, 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).

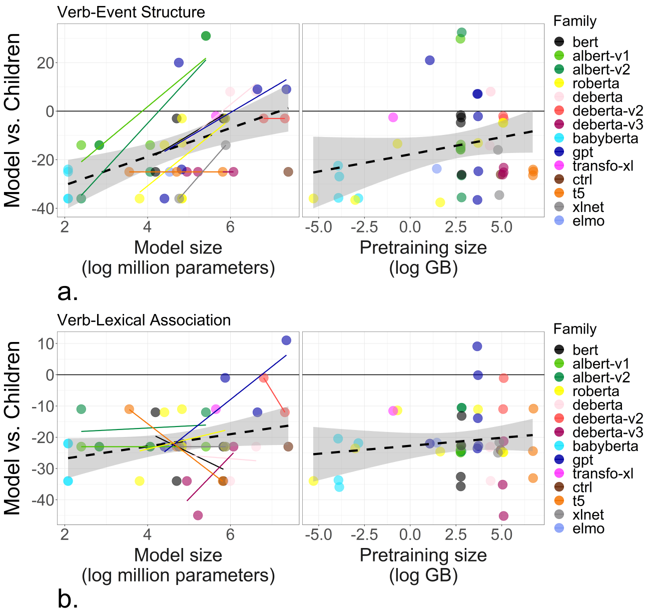

Figure 4: Models' distance from children by model size and pretraining size, when comparing verb-event vs. control (a) and verb-lexical vs. control conditions (b).

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 will examine by-item performance across models. This will help us assess how prior context associations for each target word impact models' performance, and whether this relation aligns with that observed in children.

Basing sense prototypes on child input partially tuned models' representations to children's. 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, likely not present in the adult language used for pretraining (e.g., Wikipedia). One way to align models with the child environment would be pretraining directly on child input (Hosseini et al., 2022; Warstadt & Bowman, 2022). However, this task is limited by the lack of sufficiently large corpora. For example, in our study we included BabyBERTa (Huebner, Sulem, Cynthia, & Roth, 2021), which despite being pretrained on child input showed no sensitivity to sentence context, likely due to its small pretraining (5 million tokens).

Models' performance was impaired in tasks that introduced contrastive cues (Figure 3 and 4). This suggests that previous results showing that Transformers can even approximate adult performance (Loureiro et al., 2021) still require further investigation. Sense prototypes based on child input might have contributed to the low performance of the models in our study. In upcoming work, we will examine the role of reference sentences by additionally using adult-directed sense-tagged sentences (from the spoken part of the British National Corpus; BNC Consortium, 2007). Given that previous studies have not used contrastive tasks, another possibility is that such tasks might 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. 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. 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, Goodman, & Tenenbaum, 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, Alishahi, & Chrupała, 2022). Finally, 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, Tenenbaum, & Regier, 2011).

We began to examine the capabilities and limitations of Transformer models for studying early word sense disambiguation. 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. We emphasized the importance of filling the gap between children and models by integrating multimodal knowledge in neural systems.

Acknowledgments

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).

References

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). Berlin, Heidelberg: Springer. Retrieved 2023-01-05, from https://doi.org/10.1007/978-3-642-31863-4_11 doi: 10.1007/978-3-642-31863-4_11

Ambridge, B. (2019, September). Against stored abstractions: A radical exemplar model of language acquisition:. *First Language*. Retrieved 2020-04-07, from https://journals.sagepub.com/doi/10.1177/0142723719869731 (Publisher: SAGE PublicationsSage UK: London, England) doi: 10.1177/0142723719869731

Ambridge, B. (2020, October). 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. Retrieved 2021-10-11, from https://doi.org/10.1177/0142723720949723 (Publisher: SAGE Publications Ltd) doi: 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. Retrieved 2021-05-15, from http://www.proquest.com/docview/1700661308/6829D1F540F9493DPQ/56 (Num Pages: 35)

BNC Consortium. (2007, March). *British National Corpus, XML edition.* Retrieved 2023-01-19, from https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2554 (Accepted: 2018-07-27 Artwork Medium: Digital bitstream Interview Medium: Digital bitstream Publisher: University of Oxford)

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2022, July). *On the Opportunities and Risks of Foundation Models.* arXiv. Retrieved 2023-01-04, from http://arxiv.org/abs/2108.07258 (arXiv:2108.07258 [cs])

Bybee, J. (2010). *Language, Usage and Cognition.* Cambridge: Cambridge University Press. Retrieved 2023-01-05, from https://www.cambridge.org/core/books/language-usage-and-cognition/46BF7213957AF53492A7B03A9BCE9DA0 doi: 10.1017/CBO9780511750526

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2022a, June). ChiSense-12: An English Sense-Annotated Child-Directed Speech Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5198–5205). Marseille, France: European Language Resources Association. Retrieved 2023-01-16, from 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). Retrieved 2023-01-06, from https://escholarship.org/uc/item/9kh29212

Chang, T. A., & Bergen, B. K. (2021, October). *Word Acquisition in Neural Language Models.* arXiv. Retrieved 2023-01-19, from http://arxiv.org/abs/2110.02406 (arXiv:2110.02406 [cs]) doi: 10.48550/arXiv.2110.02406

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv. Retrieved 2023-01-17, from http://arxiv.org/abs/1810.04805 (arXiv:1810.04805 [cs]) doi: 10.48550/arXiv.1810.04805

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

Haber, J., & Poesio, M. (2020, June). Word Sense Distance in Human Similarity Judgements and Contextualised Word Embeddings. In *Proceedings of the Probability and Meaning Conference (PaM 2020)* (pp. 128–145). Gothenburg: Association for Computational Linguistics. Retrieved 2023-01-16, from 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. Retrieved 2023-01-06, from https://onlinelibrary.wiley.com/doi/abs/10.1002/aur.1487 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aur.1487) doi: 10.1002/aur.1487

Hewitt, J., & Manning, C. D. (2019, June). A Structural Probe for Finding Syntax in Word Representations. In *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)* (pp. 4129–4138). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved 2021-10-13, from https://aclanthology.org/N19-1419 doi: 10.18653/v1/N19-1419

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022, October). *Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training.* bioRxiv. Retrieved 2023-01-05, from https://www.biorxiv.org/content/10.1101/2022.10.04.510681v1 (Pages: 2022.10.04.510681 Section: New Results) doi: 10.1101/2022.10.04.510681

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021, November). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 624–646). Online: Association for Computational Linguistics. Retrieved 2022-10-26, from https://aclanthology.org/2021.conll-1.49 doi: 10.18653/v1/2021.conll-1.49

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651–3657). Florence, Italy: Association for Computational Linguistics. Retrieved 2023-01-05, from https://www.aclweb.org/anthology/P19-1356 doi: 10.18653/v1/P19-1356

Khanna, M. M., & Boland, J. E. (2010, January). Children's use of language context in lexical ambiguity resolution. *Quarterly Journal of Experimental Psychology*, *63*(1), 160–193. Retrieved 2023-01-06, from https://doi.org/10.1080/17470210902866664 (Publisher: SAGE Publications) doi: 10.1080/17470210902866664

Laverghetta Jr, A., & Licato, J. (2021, April). Modeling Age of Acquisition Norms Using Transformer Networks. *The International FLAIRS Conference Proceedings*, *34*. Retrieved 2023-01-05, from https://journals.flvc.org/FLAIRS/article/view/128334 doi: 10.32473/flairs.v34i1.128334

Loureiro, D., Jorge, A. M., & Camacho-Collados, J. (2022, April). LMMS Reloaded: Transformer-based Sense Embeddings for Disambiguation and Beyond. *Artificial Intelligence*, *305*, 103661. Retrieved 2023-01-05, from http://arxiv.org/abs/2105.12449 (arXiv:2105.12449 [cs]) doi: 10.1016/j.artint.2022.103661

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021, March). *Analysis and Evaluation of Language Models for Word Sense Disambiguation.* arXiv. Retrieved 2023-01-05, from http://arxiv.org/abs/2008.11608 (arXiv:2008.11608 [cs]) doi: 10.48550/arXiv.2008.11608

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

Mani, N., Daum, M. M., & Huettig, F. (2016, November). "Proactive" in many ways: Developmental evidence for a dynamic pluralistic approach to prediction. *Quarterly Journal of Experimental Psychology*, *69*(11), 2189–2201. Retrieved 2021-04-24, from https://doi.org/10.1080/17470218.2015.1111395 (Publisher: SAGE Publications) doi: 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. (Place: US Publisher: American Psychological Association) doi: 10.1037/rev0000126

Melamud, O., Goldberger, J., & Dagan, I. (2016, August). context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning* (pp. 51–61). Berlin, Germany: Association for Computational Linguistics. Retrieved 2023-01-16, from https://aclanthology.org/K16-1006 doi: 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). Retrieved 2023-01-06, from https://escholarship.org/uc/item/1pq031fn

Nikolaus, M., Alishahi, A., & Chrupała, G. (2022, September). Learning English with Peppa Pig. *Transactions of the Association for Computational Linguistics*, *10*, 922–936. Retrieved 2023-05-08, from https://doi.org/10.1162/tacl_a_00498 doi: 10.1162/tacl_a_00498

Nikolaus, M., & Fourtassi, A. (2021, June). Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 200–210). Online: Association for Computational Linguistics. Retrieved 2023-01-19, from https://aclanthology.org/2021.cmcl-1.24 doi: 10.18653/v1/2021.cmcl-1.24

Nyamapfene, A., & Ahmad, K. (2007, August). A Multimodal Model of Child Language Acquisition at the One-Word Stage. In *2007 International Joint Conference on Neural Networks* (pp. 783–788). (ISSN: 2161-4407) doi: 10.1109/IJCNN.2007.4371057

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011, March). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338. Retrieved 2023-01-26, from https://www.sciencedirect.com/science/article/pii/S0010027710002593 doi: 10.1016/j.cognition.2010.11.001

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved 2023-01-16, from https://aclanthology.org/N18-1202 doi: 10.18653/v1/N18-1202

Pinker, S. (1989). *Learnability and Cognition.* Retrieved 2023-01-26, from https://mitpress.mit.edu/9780262660730/learnability-and-cognition/

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.. Retrieved 2023-01-17, from https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996, December). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. Retrieved 2023-01-06, from https://www.science.org/doi/abs/10.1126/science.274.5294.1926 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.274.5294.1926

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., … Fedorenko, E. (2021, November). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. Retrieved 2022-11-02, from https://pnas.org/doi/full/10.1073/pnas.2105646118 doi: 10.1073/pnas.2105646118

Srinivasan, M., & Rabagliati, H. (2021). The Implications of Polysemy for Theories of Word Learning. *Child Development Perspectives*, *15*(3), 148–153. Retrieved 2022-01-18, from https://onlinelibrary.wiley.com/doi/abs/10.1111/cdep.12411 (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdep.12411) doi: 10.1111/cdep.12411

Tenney, I., Das, D., & Pavlick, E. (2019, August). *BERT Rediscovers the Classical NLP Pipeline.* arXiv. Retrieved 2023-01-05, from http://arxiv.org/abs/1905.05950 (arXiv:1905.05950 [cs]) doi: 10.48550/arXiv.1905.05950

Thornton, R. (2012, February). Studies at the interface of child language and models of language acquisition. *First Language*, *32*(1-2), 281–297. Retrieved 2023-01-26, from https://doi.org/10.1177/0142723711403881 (Publisher: SAGE Publications Ltd) doi: 10.1177/0142723711403881

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017, December). *Attention Is All You Need.* arXiv. Retrieved 2023-01-05, from http://arxiv.org/abs/1706.03762 (arXiv:1706.03762 [cs]) doi: 10.48550/arXiv.1706.03762

Warstadt, A., & Bowman, S. R. (2022, August). *What Artificial Neural Networks Can Tell Us About Human Language Acquisition.* arXiv. Retrieved 2023-01-19, from http://arxiv.org/abs/2208.07998 (arXiv:2208.07998 [cs]) doi: 10.48550/arXiv.2208.07998

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019, October). *Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings.* arXiv. Retrieved 2023-01-17, from http://arxiv.org/abs/1909.10430 (arXiv:1909.10430 [cs]) doi: 10.48550/arXiv.1909.10430