Is Child-Directed Language Optimized for Word Learning? A Computational Study of Verb Meaning Acquisition

Francesca Padovani (f.padovani@rug.nl)
Jaap Jumelet (j.w.d.jumelet@rug.nl)
Yevgen Matusevych (yevgen.matusevych@rug.nl)
Arianna Bisazza (a.bisazza@rug.nl)
Center for Language and Cognition (CLCG), University of Groningen

Abstract

Is child-directed language (CDL) optimized to support language learning, and which aspects of linguistic development does it facilitate? We investigate this question using neural language models trained on CDL versus adult-directed language (ADL). We selectively remove syntactic or lexical co-occurrence information from the model training data, and evaluate the impact of these manipulations on verb meaning acquisition. While disrupting syntax impairs learning across all datasets, models trained on CDL and spoken ADL show significantly higher resilience than those trained on written input. Tracking semantic and syntactic performance over training, we observe a "semantic-first" trajectory, with verb meanings emerging prior to robust syntactic proficiency, an asynchrony most pronounced in the spoken domain, especially CDL. These results suggest that the advantage for verb learning previously attributed to CDL may instead reflect broader properties of the spoken register, rather than a uniquely CDL-specific optimization[1].

Keywords: syntactic bootstrapping; word learning; language models; language acquisition; child-directed language

Introduction

Child-directed language (CDL)[2] is thought to provide input that is particularly conducive to learning, potentially offering advantages over adult-directed language (ADL) (Ferguson, 1964). This perspective has influenced both theoretical and empirical studies on early language acquisition (Golinkoff et al., 2015; Schick et al., 2022). Recently, however, this view has been critically reassessed by Kempe et al. (2024), who suggest that evidence for facilitative effects of CDL is largely restricted to certain linguistic features, such as prosody and register discrimination, and may not generalize to other aspects of language (e.g., morphosyntax and semantic learning). These observations indicate that, although some facilitative effects of CDL have been demonstrated, systematic comparisons with ADL across broader linguistic features remain limited.

Since children's natural linguistic environments cannot be experimentally manipulated without ethical concerns, it remains difficult to directly assess how different characteristics of the input shape learning trajectories. Computational modeling has therefore emerged as a valuable tool for studying language development, providing a controlled setting in which theoretical hypotheses can be tested. Recent advances in Natural Language Processing make it possible to train large-scale language models (LMs) under different input conditions, allowing researchers to examine how specific properties of the input influence emerging linguistic patterns (Portelance & Jasbi, 2024; Warstadt & Bowman, 2022). Reflecting the growing interest in developmentally plausible corpora, community initiatives such as the BabyLM Challenge (Charpentier et al., 2025; Wilcox et al., 2025) have promoted systematic investigations of how training data properties influence model behavior, with much of the work to date focusing on syntax and comparing models trained on CDL to those trained on ADL (Feng et al., 2024; Huebner et al., 2021). Evidence regarding the advantages of CDL for syntactic learning is inconclusive, and recent evaluations indicate that its benefits disappear when frequency effects are controlled for (Padovani et al., 2025).

Beyond syntax, semantic development constitutes another central component of early language acquisition, and verb learning in particular poses a well-known challenge due to the abstract and relational nature of verb meaning (Gentner, 1978). This study therefore investigates whether CDL is structured to facilitate the acquisition of verb meaning. This issue has been previously examined by You et al. (2021), who argue that the statistical properties of English CDL, relative to both spoken and written ADL, support the acquisition of verb categories without reliance on high-level syntactic knowledge. On this basis, they propose that CDL may enable syntax-independent semantic inference, challenging the *syntactic bootstrapping* hypothesis (Gleitman, 1990). However, their analysis is limited both in scope, focusing only on the distinction between causative and noncausative verbs, and in coverage, as it includes only a small set of verbs for each category. Combined with the use of a non-contextual distributional model, these constraints raise questions about the generalizability of their conclusions to other verbs and to other types of statistical learners. More recently, Zhu et al. (2025), also working on English, adopt a complementary perspective by assuming a central role for syntactic bootstrapping in children, and testing whether contextual Transformer-based LMs trained on CDL exhibit similar learning behavior to humans. By manipulating the input data prior to training, they show that altering syntactic structure impairs verb meaning acquisition more than does perturbing lexical co-occurrences. These results suggest that LMs, like human learners, rely on syntactic cues to infer verb meaning. However, Zhu et al.'s analysis is limited to CDL, leaving open the question of whether this type of language

[1] Repository: https://github.com/fpadovani/cdl_verb_meaning

[2] Throughout this paper, we use the term CDL specifically to refer to transcripts of child-directed speech.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

is more robust to syntax degradation when acquiring verb meaning, compared to other input domains.

In the present study, we address this gap by explicitly comparing CDL to both spoken and written ADL, being guided by three main research questions:

- **RQ1:** Does perturbing syntactic structure affect verb meaning acquisition more than perturbing lexical co-occurrence only in CDL or also in ADL?
- **RQ2:** Is verb meaning acquisition in CDL, compared to ADL, more robust to the ablation of syntax?
- **RQ3:** How do the developmental trajectories of verb meaning acquisition and syntactic performance relate over time? How do these trajectories differ between CDL and ADL?

To address these questions, we train autoregressive Transformer LMs and use the data manipulation framework of Zhu et al. (2025), applying controlled perturbations to the training corpora (**RQ1, RQ2**). Verb meaning acquisition is evaluated with in-domain semantic minimal pairs, and its relation to syntactic development is assessed using established syntactic minimal pair benchmarks (**RQ3**).

Related Work

The Verb Learning Problem Understanding how children acquire verb meaning has long been a central problem in language acquisition: unlike nouns, verbs encode relational meaning and require learners to infer abstract event structures from sparse and noisy linguistic input (Gentner, 1978). Much work emphasizes the role of syntax in constraining verb interpretation, an idea formalized in the syntactic bootstrapping hypothesis (Babineau et al., 2024; Gleitman, 1990). Support for this view comes, among others, from Gillette et al. (1999), who show that verbs are more readily identified when learners have access to syntactic frames than to cross-situational cues alone. Other studies demonstrate that learners can extract aspects of verb meaning from contextual regularities in the input, even in the absence of fully developed syntactic representations (Newport & Aslin, 2004; Smith & Yu, 2008). Empirical evidence suggests that children flexibly integrate multiple sources of information during verb learning, including word co-occurrence statistics and syntactic frames (Lidz et al., 2003; Naigles, 1996).

Assessing Linguistic Performance in LMs Inspired by language acquisition research, recent work in Computational Linguistics has examined how lexical meaning emerges in LMs trained on developmentally plausible input, such as CHILDES transcripts (MacWhinney, 2000). A common category of evaluation, surprisal-based approaches, assesses word learning by evaluating a model's ability to anticipate words in context: words that are highly predictable in a given context elicit lower surprisal, reflecting the model's confidence in its predictions (Chang & Bergen, 2022; Portelance et al., 2023; Shafiabadi & Wisniewski, 2025). However, such measures have limitations: they show when words are predicted reliably but not which cues the model uses to learn them. Moreover, even frequent words can appear surprising in unusual contexts, since surprisal mainly reflects the next-word prediction objective rather than abstract word knowledge. Complementary to surprisal-based approaches, lexical decision paradigms probe whether models distinguish real words from plausible non-words as a proxy for semantic performance (Bunzeck & Zarrieß, 2025; Goriely et al., 2024). Minimal pairs offer another well-established paradigm for assessing linguistic knowledge in LMs. Based on the observation that small changes can render a sentence acceptable or unacceptable to native speakers (Chomsky, 2014), minimal pairs test whether a model assigns higher probability to the grammatical or contextually appropriate sentence when only a single element differs (Linzen et al., 2016; Marvin & Linzen, 2018). Benchmarks like BLiMP (Warstadt et al., 2020), Zorro (Huebner et al., 2021) and CLAMS (Mueller et al., 2020) primarily evaluate syntax, whereas semantic minimal pairs assess whether models prefer the contextually or semantically correct verb over an alternative (Zhu et al., 2025).

Computational Studies of Verb Learning Mechanisms Two recent studies are particularly inspirational for this work.

- You et al. (2021) (henceforth **Y21**) test whether distributional statistics in child-directed language suffice for verb meaning acquisition. They train Word2Vec models (Mikolov et al., 2013), non-contextual distributional models, on CDL and on spoken and written ADL. Focusing on causative versus non-causative verbs, they find that models trained on CDL distinguish these verb classes more accurately than models trained on ADL, despite not having access to explicit syntactic structure. This discriminative power is strongest in smaller context windows, suggesting that CDL's repetitive and predictable patterns provide dense, redundant cues helping learners infer semantic regularities from the input. Although limited to a small set of verbs (23 causative + 9 non-causatives), their conclusion aligns with Futrell (2025), who emphasizes the critical role of local redundancy in language learning and highlights how distributional statistics can support the learning of form–meaning mappings, even when learners only have access to surface forms. Relatedly, Lee and Berg-Kirkpatrick (2025) also show that model learnability is predicted by statistical simplicity, operationalized as reduced n-gram complexity, highlighting the importance of regular and redundant input patterns for language modeling. We adopt from Y21 the comparative setup that contrasts CDL with spoken and written ADL. Additionally, their findings motivate our hypothesis that CDL may be especially well suited to support verb learning even when syntactic cues are limited.
- Zhu et al. (2025) (henceforth **Z25**) examine whether LMs exhibit syntactic bootstrapping by evaluating verb meaning acquisition across all verb forms in the training data, using both RoBERTa (Liu et al., 2019), a masked LM, and GPT-2 (Radford et al., 2019), a causal LM. Through targeted manipulations of word order and lexical co-occurrence in CDL, Z25 shows that disrupting syntactic cues leads to larger impairments in verb

learning than does perturbing distributional statistics, providing evidence for syntactic bootstrapping in LMs trained on this domain. Because their analysis is restricted to CDL, it remains unclear how the relative importance of syntactic cues varies across CDL and spoken and written ADL. We adopt from Z25 the methodological framework involving the selective ablation of syntactic and co-occurrence information, to systematically probe verb meaning acquisition across different input domains.

Experimental Setup

We assess verb meaning acquisition in English by comparing Tranformer-based LMs trained on CDL to those trained on three additional corpora. Following Z25, we train models on both original and manipulated versions of each corpus. Verb meaning acquisition is evaluated via in-domain semantic minimal pairs for each dataset. Finally, we examine how verb learning relates to the emergence of syntactic knowledge by testing models on standard syntactic minimal pair benchmarks (Huebner et al., 2021; Padovani et al., 2025; Warstadt et al., 2020).

Training Data

We follow Y21 in using CDL and the spoken and written portions of BNC (Davies, 2004), while also including Wikipedia to provide a complementary written domain with encyclopedic prose, extending the analysis beyond the mostly narrative content of the BNC. We base the size of all datasets on the amount of data available in the CDL corpus, which, after preprocessing, comprises a total of 15M tokens (10/2.5/2.5M for training/dev/test set, respectively). All other datasets are split according to the same train/dev/test proportions.

Child-Directed Language Following Z25, we use developmentally plausible CDL data from *CHILDES* (MacWhinney, 2000), specifically the CHILDES-only portion of the BabyLM Challenge corpus[3] (Charpentier et al., 2025). The raw corpus contains roughly 29M tokens; after removing speaker tags and paralinguistic material, 15M tokens remain.

ADL Written – British National Corpus We extract a portion of the written part of the *British National Corpus* (BNC) (Davies, 2004). The BNC comprises 100M words spanning a wide range of genres, including spoken language, fiction, magazines, newspapers, and academic texts. We analyze the distribution of documents in the written portion, observing a predominance of books, periodicals, and TV news.

ADL Written – Wikipedia *Wikipedia* is a collection of written articles that have an encyclopedic format, and it is commonly used as a standard resource for training LMs. We use the existing Wikipedia splits from Huebner et al. (2021).[4]

[3]https://babylm.github.io/

[4]https://github.com/phueb/BabyBERTa

Table 1: Descriptive statistics of the training datasets. TTR refers to type-token ratio.

Statistic	CDL	CANDOR	BNC	Wikipedia
1-gram TTR	0.004	0.005	0.013	0.021
2-gram TTR	0.109	0.124	0.269	0.309
3-gram TTR	0.420	0.442	0.684	0.715
Avg sent length	4.55	9.31	20.40	20.34

ADL Spoken – CANDOR + Switchboard + BNC To represent adult conversational language, we construct a spoken-domain dataset using *CANDOR*[5] (Reece et al., 2023), collected from video-call recordings between strangers during the 2020 pandemic. As CANDOR contains fewer than 15M tokens, we supplement it with the *Switchboard Dialog Act Corpus* (Stolcke et al., 2000) and a sample from the spoken portion of the *BNC* (Davies, 2004), achieving the target dataset size.

Table 1 summarizes key statistics of the training datasets. As expected, the written corpora exhibit higher type-token ratios and longer average sentence lengths than the spoken datasets, reflecting greater lexical diversity and syntactic complexity.

Training Data Manipulations

To investigate how syntactic and lexical co-occurrence information contribute to verb meaning acquisition, we apply targeted perturbations to the training data following the framework of Z25.[6] We then compare the effects of these manipulations across CDL and ADL to assess whether the patterns observed in CDL generalize or diverge in other domains. Below, we describe both conditions in detail:

REPLACE.WORD – co-occurrence disruption Under this manipulation, lexical items that co-occur with verbs, specifically nouns, adjectives, and adverbs, are replaced within each sentence by other words with the same part-of-speech and fine-grained XPOS tag (i.e., language-specific morphological tags encoding features such as number, tense, or degree). Replacement words are sampled according to frequency bins for the corresponding POS and XPOS tags, thus preserving the overall token frequency distribution. In sentences containing multiple verbs, the root verb is left unchanged, while dependent verbs are replaced following the same procedure as for other parts of speech. This manipulation preserves surface word order and syntactic structure while selectively removing verb-specific distributional cues. The proportion of replaced words is 10.3% for CDL, 11.3% for CANDOR, 26.3% for Wikipedia, and 24.2% for BNC, reflecting differences in average sentence length across domains (Table 1). This manipulation contributes to **RQ1** and is used in combination with SHUFFLE.ORDER to test whether the relative impact of

[5]https://guscooney.com/candor-dataset/

[6]The original Z25 code was not publicly available at the time of this work, so we independently implemented the perturbation pipeline.

co-occurrence versus syntactic disruptions observed in CDL generalizes to ADL.

SHUFFLE.ORDER – word order disruption This simple manipulation exploits the fact that English encodes syntactic structure primarily through word order. Words are randomly shuffled within each sentence, eliminating linear order dependencies and degrading syntactic information while retaining sentence-level co-occurrence statistics. This manipulation directly operationalizes our **RQ2**: by disrupting word order within sentences in CDL vs. ADL we can assess whether verb meaning acquisition is more resilient in CDL.

Models and Training

We focus on autoregressive LMs, specifically GPT-2 (Radford et al., 2019), due to the higher cognitive plausibility of their training objective compared to bidirectional models. We follow the setup of Z25, using a model architecture consisting of 12 layers, 12 attention heads per layer, and 768 hidden units. Training is conducted with a linear learning rate scheduler, a learning rate of 10^{-4}, and a batch size of 256. To examine the joint development of semantic and syntactic performance over an extended learning period (**RQ3**), models are trained for 20 epochs. For each dataset and experimental condition, results are averaged across five random initializations (seeds). Results reported for **RQ1** and **RQ2** are based on the checkpoint with the lowest validation loss on the development set for each corpus and experimental condition, as mild overfitting was typically observed after several training epochs.

Evaluation

We adopt a minimal pair paradigm commonly used to probe linguistic knowledge in LMs. Each test item consists of a pair of sentences that differ only in one targeted property, while all other aspects are held constant. A model is considered to make a correct prediction if it assigns a higher probability to the grammatical or contextually appropriate sentence than to its minimally altered counterpart. Sentence probabilities are computed using the `minicons` library (Misra, 2022).

Semantic Minimal Pairs Following Z25, we generate in-domain semantic minimal pairs from the test portion of each dataset. Since we focus on verb meaning acquisition, we identify the target verb (i.e., the root verb) in each sentence and create up to five alternative sentences by substituting verbs from the same fine-grained frequency bin, defined over POS and XPOS categories from the training data. To ensure comparability across domains, we limit sentence length to 10–30 tokens. This results in minimal pairs that differ only in the target verb, isolating the contribution of lexical–semantic appropriateness.[7] In total, we obtain four distinct sets, one per training corpus, each containing 140,000 minimal pairs and spanning a broad range of verb types (between 801 and 2,765 unique verb lemmas across datasets).

An example is provided in (1).

(1) a. You can sit out here by me on the other stool.
b. * You can *try* out here by me on the other stool.

As a sanity check, we verify that our semantic minimal pairs capture genuine verb meaning knowledge rather than dataset-specific artifacts. We use minimal pairs generated from one dataset (e.g., CDL) to test models trained on another (e.g., BNC). Models achieve high accuracy only when training and test domains match, showing that strong in-domain performance reflects true acquisition of domain-specific verb patterns rather than quirks of the evaluation procedure. All results reported in the Results section refer to in-domain train/test setup.

Syntactic Minimal Pairs To test models across a range of syntactic phenomena, we first consider two benchmarks commonly used for English LMs: BLiMP (Warstadt et al., 2020) and Zorro (Huebner et al., 2021). However, Padovani et al. (2025) recently highlighted the limitations of these benchmarks when applied to language with distinct lexical distributions or domain-specific vocabularies, such as CDL. As a solution, they proposed to generate corpus-specific test sets, ensuring the minimal pairs contain words drawn from comparable frequency ranges in each corpus. They applied this method to CLAMS (Mueller et al., 2020), resulting in the FIT-CLAMS benchmark, which we adopt for our analysis. CLAMS comprises only the syntactic paradigm of subject–verb number agreement, tested across different clause types (e.g., simple clauses, prepositional phrases, coordinates, and relative clauses). As the original FIT-CLAMS versions were developed for CDL and Wikipedia, here we also apply the same procedure to generate two new versions, for BNC and CANDOR. This ensures that syntactic performance can be probed under comparable conditions across all four datasets. Example (2) illustrates a syntactic minimal pair testing *subject–verb number agreement*:

(2) a. The painter in front of the waiter enjoys.
b. * The *painters* in front of the waiter enjoys.

In the next section, we report results for **RQ1** and **RQ2**, evaluating models trained under different data manipulations on semantic minimal pairs, and for **RQ3**, tracking accuracy on semantic and syntactic minimal pairs over time for models trained on the original datasets.

Results

RQ1: Effect of Co-Occurrence Disruptions on Verb Learning We first replicate Z25's finding that models trained on CDL are more sensitive to syntactic disruptions than to changes in lexical co-occurrence when evaluated on semantic minimal pairs targeting verb meaning. Perturbing word order produces a larger drop in accuracy than does altering verb co-occurrences, suggesting that LMs, like human learners, rely on syntactic

[7] Two methodological considerations apply: first, some frequency bins are sparse and do not always contain five suitable replacement verbs, leading to fewer than five generated minimal pairs. Second, this procedure does not guarantee that every alternative is strictly unacceptable; nevertheless, substituting the target verb with a contextually mismatched alternative should produce less plausible sentences on average, even if not at the individual level.

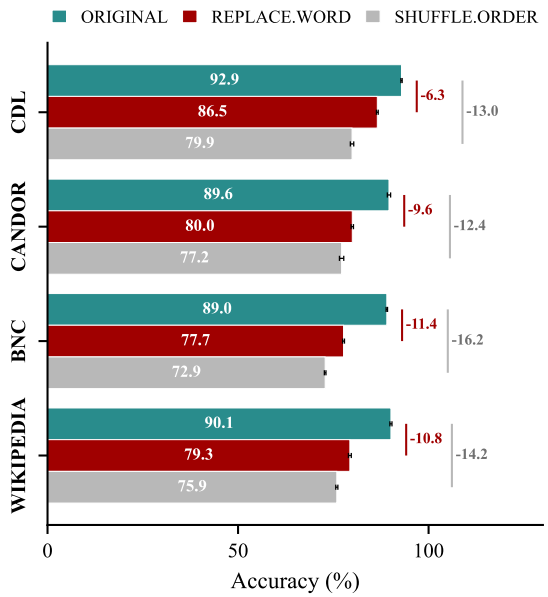

Figure 1: Accuracy scores on semantic minimal pairs in the three different experimental conditions for each data domain.

scaffolding to infer verb meaning. We then extend this analysis to the other three corpora to examine whether the relative impact of syntactic versus lexical disruptions observed in CDL generalizes to adult-directed input. As shown in Figure 1, the accuracy drop from ORIGINAL to REPLACE.WORD (red) is consistently smaller than the drop from ORIGINAL to SHUFFLE.ORDER (gray), not only in CDL but also in spoken (CANDOR) and written ADL (BNC, Wikipedia). While this differential impact is shared across domains, the magnitude of the drop caused by lexical perturbations varies with corpus properties. In particular, the accuracy drop under this manipulation scales with average sentence length (reported in Table 1 above) and with the lexical replacement rates as reported in the REPLACE.WORD manipulation section. Intuitively, longer sentences with more words co-occurring with the main verb lead to larger drops under lexical perturbations. Overall, these results answer **RQ1** by showing that syntactic cues provide critical scaffolding for verb learning in both CDL and ADL, while lexical co-occurrences exert a more limited influence.

RQ2: Effect of Syntax Perturbation on Verb Learning **RQ2** directly builds on the results of **RQ1** and on the observation that syntax plays a key role in verb meaning acquisition across all data domains. Drawing on Y21's claim that CDL may be optimized for syntax-free semantic inference, we ask whether the impact of syntactic perturbations differs across domains and whether CDL shows greater resilience to word order disruptions than written or spoken ADL. While we expect CDL to maintain relatively high accuracy due to its repetitive and predictable input, we do not make strong predictions for adult-directed spoken (CANDOR) or written (Wikipedia, BNC) corpora, although the greater syntactic complexity of written text may make it more sensitive to word order manipulations. To formally evaluate the robustness of verb learning when syntax is ablated, we employed an Ordinary Least Squares (OLS) regression model[8] to predict accuracy as a function of Dataset (CDL, CANDOR, BNC, Wikipedia), Condition (ORIGINAL vs. SHUFFLE.ORDER), and their interaction. By setting the CDL baseline in the ORIGINAL condition as our reference point (the intercept), we could precisely quantify the performance decrement associated with syntactic disruption and determine whether the drop in accuracy varied significantly by corpus. As summarized in Figure 1, models trained on CDL achieve high baseline performance (92.9%), which decreases by 13.0% when word order is corrupted, a significant main effect of Condition ($\beta = -0.13$, $p < .001$). Crucially, the interaction terms reveal that the impact of syntactic ablation is not uniform across data domains. The accuracy drop is significantly more pronounced in the written corpora than in CDL, with BNC and Wikipedia showing additional decreases of 3.2% ($\beta = -0.032$, $p < .001$) and 1.2% ($\beta = -0.012$, $p < .001$), respectively. By contrast, the interaction term for the adult-directed spoken corpus (CANDOR) is not significant ($\beta = 0.005$, $p = .088$). These results suggest that the robustness to syntactic perturbation is not exclusive to CDL. Rather, the comparable performance of CANDOR indicates that a reduced reliance on word order may be a general characteristic of learning from conversational speech, contrasting with the higher sensitivity to structural changes in representations derived from written text.

RQ3: Developmental Trajectories of Semantics and Syntax To address **RQ3**, we examine the evolution of models' performances over the course of training, comparing the emergence of syntactic proficiency with the acquisition of verb meaning. Our hypothesis is that the distributional properties of CDL allow models to form robust representations of verb meaning early in training, even before a sharp increase in syntactic knowledge is observed. In contrast, we expect that in more complex adult-directed domains, verb meaning development depends more strongly on—and thus progresses alongside—emerging syntactic proficiency. We focus our analysis on the FIT-CLAMS benchmark, which provides the most reliable cross-domain comparisons. However, we also test our models on BLiMP and Zorro, finding consistent overall patterns. The learning trajectories shown in Figure 2 reveal a "semantic-first" trend across all domains: Semantic Accuracy (middle panel) rises steadily in early epochs (0.063 to 1), while Syntactic Accuracy remains near chance. However, the datasets differ in the degree of synchrony between these two competencies. This is explicitly captured by the Ratio Sem/Syn (bottom panel), where a higher ratio indicates a greater developmental lead for semantics over syntax. CDL demonstrates the most pronounced developmental asynchrony;

[8] We initially tested a mixed-effects model with seed as a random effect, but the random effect variance was effectively zero, so we used a standard linear regression instead.

it maintains the highest ratio of semantic-to-syntactic knowledge from Epoch 0.04 through to the final checkpoint. By Epoch 1 (vertical line in Figure 2), CDL reaches almost 80% semantic accuracy, while syntactic performance remains under 60%. By contrast, written ADL (Wikipedia and BNC) shows a much tighter coupling of these skills. After a brief early semantic lead, the ratio values for these two datasets drop steeply after Epoch 3 as syntactic proficiency rapidly "catches up" to semantic levels, resulting in a more balanced final profile. Interestingly, CANDOR occupies a middle ground, highlighting the distinction between spoken and written input. While CANDOR aligns with the CDL trajectory early on, showing a similarly high semantic lead, it eventually diverges as its syntactic acquisition accelerates, landing between the "coupled" written corpora and the "decoupled" CDL. Overall, we observe that different training domains are associated with distinct developmental trajectories. Throughout the learning process, CDL consistently maintains the highest ratio of semantic-to-syntactic knowledge. While this trend aligns with our hypothesis that CDL is optimized for verb meaning that relies less on formal syntactic proficiency, we remain cautious about drawing causal conclusions. These observations remain tied to our specific operationalization of syntactic acquisition, a limitation we discuss below.

Discussion and Conclusion

Across all domains, our results indicate that verb meaning acquisition depends more strongly on syntactic structure than on lexical co-occurrence. Importantly, the relative resilience to syntactic disruption observed for CDL does not appear to be unique to child-directed input: models trained on conversational language, including both CDL and spoken ADL (CANDOR), exhibit greater stability than those trained on written text. This pattern suggests that differences in robustness to syntactic perturbations are better explained by spoken versus written input, rather than CDL versus ADL specifically, providing a more nuanced perspective compared to You et al. (2021), who emphasize CDL as uniquely optimized for verb learning. Longitudinally, we observe a "semantic-first" trajectory where verb meanings emerge before robust syntactic proficiency across all domains, but with systematic differences in how tightly semantic and syntactic development become coupled over time. This developmental asynchrony is most pronounced within the spoken domain, particularly in CDL. Together, these findings suggest that, while syntax serves as a universal scaffold for verb learning, conversational input—whether child- or adult-directed—may support more robust semantic representations even when syntactic knowledge remains limited.

Nevertheless, our study is limited by its coarse-grained approach to syntax, treating it primarily as word-order preservation within an English-centric context. This coarse-grained approach leaves open which specific syntactic frames or cues drive verb learning—for instance, whether transitive verbs require only correct agent–patient ordering or additional contextual signals. Extending these manipulations, e.g., by selectively scrambling different constituents or preserving local hierarchical structures, would enable a more fine-grained causal investigation of the sources of syntactic bootstrapping in language models.

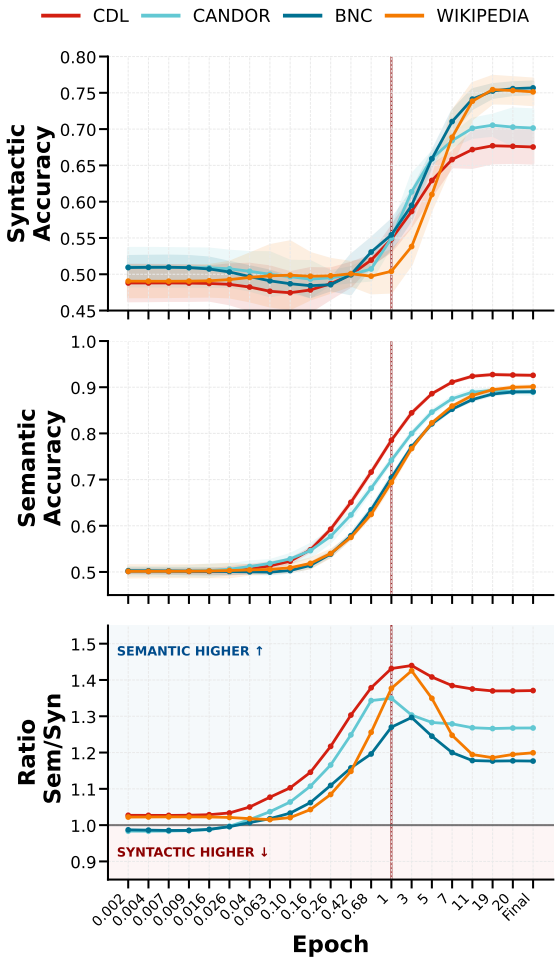

Figure 2: Evolution of model performance across training checkpoints for the four datasets considered in this study. Top: syntactic accuracy on FIT-CLAMS. Middle: semantic accuracy on semantic minimal pairs. Bottom: ratio of semantic to syntactic accuracy.

Methodologically, tracking semantic and syntactic trajectories opens avenues for analyses beyond behavioral probing. Future work could integrate interpretability methods, e.g., monitoring the emergence of Syntactic Attention Structure (SAS) (Chen et al., 2024), as a complementary assessment of syntactic knowledge. Utilizing such model internals could help overcome limitations of minimal pair benchmarks by providing a more direct view of how structural dependencies are encoded and emerge over time. Finally, extending this framework to typologically diverse languages is crucial to determine whether the observed patterns reflect universal properties or are specific to English.

Acknowledgements

This publication is part of the project 'Polyglot Machines' (VI.Vidi.221C.009) funded by the Talent Programme of the Dutch Research Council (NWO). The authors thank the members of the InCLow and GroNLP groups at the University of Groningen for their useful feedback.

References

Babineau, M., Barbir, M., de Carvalho, A., Havron, N., Dautriche, I., & Christophe, A. (2024). Syntactic bootstrapping as a mechanism for language learning. *Nature Reviews Psychology*, *3*(7), 463–474.

Bunzeck, B., & Zarrieß, S. (2025). Subword models struggle with word learning, but surprisal hides it. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 286–300). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-short.24

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, *10*, 1–16. https://doi.org/10.1162/tacl_a_00444

Charpentier, L., Choshen, L., Cotterell, R., Gul, M. O., Hu, M. Y., Liu, J., Jumelet, J., Linzen, T., Mueller, A., Ross, C., Shah, R. S., Warstadt, A., Wilcox, E. G., & Williams, A. (Eds.). (2025). *Proceedings of the First BabyLM Workshop*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.babylm-main.0

Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., & Saphra, N. (2024). Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=MO5PiKHELW

Chomsky, N. (2014). *Aspects of the theory of syntax*. MIT press.

Davies, M. (2004). BYU-BNC. *(Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc*.

Feng, S. Y., Goodman, N. D., & Frank, M. C. (2024). Is child-directed speech effective training data for language models? In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 22055–22071). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1231

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist*, *66*, 103–114.

Futrell, R. (2025). Language Learning as Codebreaking: The Key Roles of Redundancy and Locality. In C. J. Anderson, F. Mailhot, & G. Prasad (Eds.), *Proceedings of the Society for Computation in Linguistics 2025* (pp. 54–63). Association for Computational Linguistics. https://aclanthology.org/2025.scil-1.5/

Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. *Child Development*, 988–998.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)Talk to Me: The Social Context of Infant-Directed Speech and Its Effects on Early Language Acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344. https://doi.org/10.1177/0963721415595345

Goriely, Z., Diehl Martinez, R., Caines, A., Buttery, P., & Beinborn, L. (2024). From babble to words: Pre-training language models on continuous streams of phonemes. In M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt, & E. G. Wilcox (Eds.), *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning* (pp. 37–53). Association for Computational Linguistics. https://aclanthology.org/2024.conll-babylm.4/

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In A. Bisazza & O. Abend (Eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning* (pp. 624–646). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.conll-1.49

Kempe, V., Ota, M., & Schaeffler, S. (2024). Does child-directed speech facilitate language development in all domains? A study space analysis of the existing evidence. *Developmental Review*, *72*, 101121.

Lee, I., & Berg-Kirkpatrick, T. (2025). Readability learnability: Rethinking the role of simplicity in training small language models. *Conference on Language Modeling (COLM)*.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, *89*(3), 295–303.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535. https://doi.org/10.1162/tacl_a_00115

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 preprint arXiv:1907.11692*.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*

(pp. 1192–1202). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1151

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781

Misra, K. (2022). Minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Mueller, A., Nicolai, G., Petrou-Zeniou, P., Talmina, N., & Linzen, T. (2020). Cross-linguistic syntactic evaluation of word prediction models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5523–5539). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.490

Naigles, L. R. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition*, *58*(2), 221–251.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Padovani, F., Jumelet, J., Matusevych, Y., & Bisazza, A. (2025). Child-directed language does not consistently boost syntax learning in language models. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 19746–19767). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.999

Portelance, E., Duan, Y., Frank, M. C., & Lupyan, G. (2023). Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal. *Cognitive Science*, *47*(9), e13334.

Portelance, E., & Jasbi, M. (2024). The roles of neural networks in language acquisition. *Language and Linguistics Compass*, *18*(6), e70001. https://doi.org/https://doi.org/10.1111/lnc3.70001

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A., & Marin, S. (2023). The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, *9*(13), eadf3197.

Schick, J., Fryns, C., Wegdell, F., Laporte, M., Zuberbühler, K., van Schaik, C. P., Townsend, S. W., & Stoll, S. (2022). The function and evolution of child-directed communication. *PLOS Biology*, *20*(5), 1–17. https://doi.org/10.1371/journal.pbio.3001630

Shafiabadi, N., & Wisniewski, G. (2025). Beyond surprisal: A dual metric framework for lexical skill acquisition in LLMs. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 6636–6641). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.443/

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–374.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17–60). CRC Press.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. https://aclanthology.org/2020.tacl-1.25/

Wilcox, E. G., Hu, M. Y., Mueller, A., Warstadt, A., Choshen, L., Zhuang, C., Williams, A., Cotterell, R., & Linzen, T. (2025). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, *144*, 104650. https://doi.org/https://doi.org/10.1016/j.jml.2025.104650

You, G., Bickel, B., Daum, M. M., & Stoll, S. (2021). Child-directed speech is optimized for syntax-free semantic inference. *Scientific Reports*, *11*(1), 16527.

Zhu, X., McCoy, R. T., & Frank, R. (2025). The structural sources of verb meaning revisited: Large language models display syntactic bootstrapping. *arXiv preprint arXiv:2508.12482*.