Acquisition in Babies and Machines: Comparing the Learning Trajectories of LMs in Terms of Syntactic Structures (ATTracTSS Test Set)

Sarah Rossi[1,2,*,†], Guido Formichi[1,2,†], Sofia Neri[1,2,†], Tommaso Sgrizzi[1,2,†], Asya Zanollo[1,2,†], Veronica Bressan[1,3,†] and Cristiano Chesi[1,2,†]

[1] *NeTS Lab, IUSS Pavia, P.zza Vittoria 15, 27100, Pavia, Italy*
[2] *IUSS Pavia, P.zza Vittoria 15, 27100, Pavia, Italy*
[3] *Department of Linguistics and Comparative Cultural Studies, Ca' Foscari University of Venice, Fondamenta Tofetti 1075, 30123 Venice, Italy*

Abstract

A cognitively plausible language model should (i) process language incrementally, (ii) be trained on naturalistic input, and (iii) mirror the developmental stages observed in child language acquisition. This study focuses on the third point by exploring the adherence of language models' developmental patterns to the predictions of two empirically grounded theories of syntactic acquisition, the Growing Trees and the Neo-Emergentist approaches. Using an evaluation method based on perplexity, we test whether small and medium Italian-tuned LMs (two small GPT2 LMs, GePpeTto, and Minerva-7B) show sensitivity to syntactic phenomena corresponding to three acquisitional stages documented in child Italian. Our results suggest that smaller open models only partially reflect the stagewise progression observed in children.

Keywords

Language acquisition, LMs, syntax, cognitive plausibility

1. Introduction

State-of-the-art Large Language Models (LLMs) demonstrate remarkable success on various linguistic benchmarks ([1], *inter alia*). However, from a linguistic perspective, they remain uninteresting from the point of view of their cognitive plausibility. In fact, their architecture and learning dynamics differ fundamentally from those of human learners, raising doubts about their relevance to linguistic inquiry [2, 3].

Nonetheless, following [4], we argue that language modeling—despite often being overlooked in theoretical linguistics—can contribute meaningfully to linguistic inquiry, provided that certain conditions on the cognitive plausibility of the model are met.

A language model (LM) that aspires to linguistic cognitive plausibility should meet at least three key criteria. First, it should process linguistic input incrementally, reflecting the word-by-word, real-time parsing observed in human sentence production and comprehension [4]. Second, it should be exposed to naturalistic training input, approximating the kind and distribution of linguistic data encountered by human learners (PoS argument, [5]). Third—and this is the focus of the present study—it should reproduce the developmental trajectory observed in first language acquisition, where syntactic competence follows structured and empirically documented stages. In line with this, we investigate whether LMs exhibit cognitive plausibility with respect to syntax by examining whether they reflect insights from linguistic theory on how humans acquire and process syntactic knowledge.

We compare two prominent approaches to syntactic development: the Growing Trees approach (GT) [6] and the Neo-Emergentist approach (NE) [7]. We argue that explicit, theoretically informed, and empirically grounded theories of language acquisition can serve as effective testing grounds for the evaluation of linguistic plausibility of LMs.

We propose an effective method for evaluating the acquisition stages reflected in various (L)LMs by collecting their perplexity estimates for sentences corresponding to stages observed in typical Italian first language development [8]. For our set of experiments, we drew from both

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[*] Corresponding author.

[†] These authors contributed equally.

sarah.rossi@iusspavia.it (S. Rossi); guido.formichi@iusspavia.it (G. Formichi); sofia.neri@iusspavia.it (S. Neri); tommaso.sgrizzi@iusspavia.it (T. Sgrizzi); asya.zanollo@iusspavia.it (A. Zanollo); veronica.bressan@unive.it (V. Bressan); cristiano.chesi@iusspavia.it (C. Chesi)
https://rossisarah.github.io/ (S. Rossi); https://www.iusspavia.it/en/contacts/guido-formichi (G. Formichi); https://www.iusspavia.it/it/rubrica/sofia-neri (S. Neri); https://tomsgrizzi.github.io/ (T. Sgrizzi); https://www.iusspavia.it/it/rubrica/asya-zanollo/ (A. Zanollo); https://www.unive.it/data/people/28977262 (V. Bressan); https://github.com/cristianochesi (C. Chesi)
0009-0007-2525-2457 (S. Rossi); 0009-0007-1307-194X (G. Formichi); 0009-0003-5456-0556 (S. Neri); 0000-0003-1375-1359 (T. Sgrizzi); 0009-0001-3987-4843 (A. Zanollo); 0000-0003-3072-7967 (V. Bressan); 0000-0003-1935-1348 (C. Chesi)

GT and NE literature to identify seventeen core phenomena, each represented by a prototypical structural pattern. To enrich the dataset, we introduced variations to these structures—e.g., changes in verbal class—resulting in a total of 90 subphenomena. For each subphenomenon, we generated 100 lexically neutral instances, yielding a comprehensive evaluation battery of 9,000 items. We tested three GPT2 small Italian LMs, ita-baseline-small and NeTS-3M [9, 10], GePpeTto—117M parameters [11]—and a larger one—Miverva-7B-base, 7B parameters [12]. Results show that Italian language models exhibit a stagewise syntactic learning trajectory that aligns more closely with the GT approach, which proves more predictive than the NE framework. We conclude that while key asymmetries remain, models trained on a minimal amount of input consisting solely of child-directed speech (e.g., NeTS-3M) can approximate the developmental patterns observed in human language acquisition.

2. Poverty of Stimulus, LLMs, and language theories

A striking difference between LLMs and natural language acquisition lies in the quantity of training data needed to achieve adult competence. A robust cross-linguistic observation in first language acquisition is that children converge on the adult grammar within a remarkably short developmental window—by approximately age 4 to 6—regardless of the language they are exposed to [13, 14], and with limited exposure to primary linguistic data, as emphasized in the Poverty of Stimulus (PoS) argument [15].

However, the PoS argument has been recently challenged by scholars who argue that LLMs represent the most empirically grounded models of language currently available, and that core features of human linguistic competence (e.g., recursion, logical inference, and hierarchical syntactic structure) may emerge spontaneously in predictive models trained on unannotated language data [16, 17]. From this perspective, LLMs question the necessity of domain-specific innate mechanisms posited by Generative Grammar (GG), suggesting that rich linguistic generalizations may arise from data-driven statistical learning, given domain-general cognitive inductive biases embedded in the artificial neural network architecture [18].

However, the debate concerns not only whether LLMs exhibit linguistic capacities or the amount of data required, but also whether they can inform a theory that accounts for the cognitive underpinnings of natural language. The issue should be addressed from multiple perspectives: by developing fine-grained performance metrics, creating relevant tasks and benchmarks, paying attention to the amount of data, and considering model architectures that may embed relevant linguistic intuitions in the form of inductive biases [4].

Within a linguistic and cognitive perspective, the distinction between models and theories is well established and relevant to discussions about LLMs. [15] distinguishes models, tools for simulating or predicting linguistic data, from theories which, on the other hand, seek to explain underlying cognitive mechanisms. [19] similarly emphasizes that valid theories must provide mechanistic explanations rather than merely replicate behavior. More recently, [20] formalizes this distinction, describing models as devices for representing systems or testing specific hypotheses, whereas theories aim to provide explanatory frameworks to generalize across phenomena. They argue that during early theory development, when empirical testing is limited, plausibility—shaped by factors such as computational tractability and theoretical invariance—serves a critical criterion for advancing from models to theories. In sum, while LLMs demonstrate impressive empirical performance and offer valuable tools for exploring linguistic patterns, their fundamental differences from human cognition, limitations in capturing graded acceptability, and reliance on vast datasets, distinguish them from genuine linguistic theories. At present, LLMs function as tools for hypothesis testing rather than as explanatory accounts of language cognitive foundations.

In this context, we draw on two linguistic theories from the current literature to support the view that language learning by (L)LMs can be meaningfully assessed—and compared to child language acquisition—using precise linguistic criteria.

3. Theories of Language Acquisition

As already mentioned, children converge on adult grammar within a remarkably short time [13, 14]. While there can be (moderate) variability in the timing of acquisition in typically developing children, the developmental patterns are consistent across individuals, in two key respects. First, all children go through stages in which they make systematic, non-random errors—such as overproduction of computationally lighter structures with a smaller number of morphosyntactic elements [21], like uninflected verbs (infinitives [22, 23] and imperatives [24]; e.g., *Mangi-a!* 'Eat!, imperative' vs. *Mangi-a-v-ano* 'They ate', past imperfective).

Second, children produce and master certain sentence types before others, and—crucially—the order of acquisition appears to be consistent across learners: some children progress more rapidly than others, but all pass through the same developmental stages. This provides further evidence that the human language faculty constrains the hypothesis space available to learners. This study focuses on this second dimension of acquisition: the order in which different sentence types are acquired.

3.1. Comparing Competing Theories

We examine two prominent theories in the literature concerning the order in which syntactic structures are acquired by children. Both seek to answer the same core question (namely, which structures emerge earlier or later in child language), but they differ significantly in their empirical methodologies and theoretical assumptions, leading to divergent predictions that remain under active investigation. Given the ongoing nature of this debate, we consider both approaches in our analysis, without prematurely excluding either.

3.2. Growing Trees Approach

The GT approach takes the syntactic tree—a symbolic and highly formalized representation of sentence structure—as its central object of study. Syntactic trees capture the hierarchical relationships among constituents, making explicit distinctions that are not evident in surface word order, and are therefore indispensable for modeling core properties of natural language.

The GT hypothesis proposes that syntactic development unfolds in a layered fashion, reflecting the gradual availability of different regions of the tree. Initially, only low structural domains, such as the verb phrase (vP) and inflectional phrase (IP), are accessible to the child, allowing for simple subject–verb sentences, for instance. Subsequently, portions of the so-called Left Periphery [? 25], a high functional layer, become available, supporting the production of wh-questions and preposed adverbs. Only later does the full functional spine, including higher CP-level structures like embedded clauses, relatives, and "why"-questions, become active. The GT model builds upon earlier maturational analyses introduced in the 1990s, notably [26], and further developed in subsequent work (see [6, 27]). In a cognitively plausible model, one would expect to observe a learning trajectory mirroring that of human acquirers, in which early-acquired structures (e.g., simple S–V sentences) are mastered before later-acquired ones (e.g., embedded clauses).

Traditional metrics for assessing language development, such as Mean Length of Utterance in words (MLUw) alone or average age of acquisition across child samples, have limited explanatory power due to the documented high degree of individual variability in acquisition speed [6]. In other words, some children are faster than others, but all of them follow the same developmental path, in that they all acquire various syntactic structures in the same order.

Empirical studies across multiple languages have shown that acquisition proceeds in structural bursts or "explosions": at a given point, an entire syntactic domain (e.g., the vP+TP layer) becomes accessible, and all structures associated with that domain become available to the child. Crucially, within these domains, there is no robust evidence for a fixed internal acquisition order, suggesting that what is developmentally primary is the availability of the domain itself, not the sequential mastery of its substructures. These domains are straightforwardly captured by the detailed cartographic structure of the functional spine as it has been drawn by theoretical linguists over the past 30 years [28, 25].

While the foundational empirical work focused on Hebrew, the GT framework has since been extended to other languages throught both experimental and corpus-based studies, including Italian [29, 30, 31], English [32], and others [27].

3.3. Neo-Emergentist Approach

The Neo-Emergentis approach [7, 33, 34] to language acquisition departs radically from both traditional nativist and certain usage-based models. This approach is theoretically motivated to a maximally impoverished Universal Grammar (UG), in line with Chomskyan "Three Factors" [35]. Rather than positing rich, innate linguistic content (Factor 1), this model shifts explanatory weight onto the interaction between primary linguistic data (PLD; Factor 2) and general cognitive learning principles (Factor 3), thereby advancing a minimalist conception of UG.

The central claim is that syntactic categories are not innately specified but are emergent, and that acquisition proceeds along a learning path where coarser-grained categories are acquired before finer-grained refinements. This involves a successive division algorithm, where the child initially makes basic contrasts (such as predicate/argument) followed by more fine-grained subdivision (identifying discourse and thematic domain up to cartographically defined syntactic distinctions). Data from Catalan, Spanish, Italian, German, and Dutch [33] suggests that basic CP structures (such as wh-questions, V2 word order, illocutionary complementisers, and topicalisation) emerge at early developmental stages (defined in terms of MLUw), challenging models that assume a fixed, innately specified hierarchy of syntactic categories [6, 26, 36]. In contrast, finer-grained structures (e.g. recursive topics, multiple left-peripheral elements, V3 orders) seem to appear only later (around or after MLUw 2.5). Crucially, building on the Peripheral Speaker-Hearer Hypothesis (PSHH), which posits that speaker-hearer perspective is formally encoded at the edges of phasal domains [37], NE model predicts that here-and-now and speaker-hearer-oriented material functions as key bootstrapping heuristics in acquisition, and therefore they are expected to be

Table 1
Stage development predictions of the two approaches. A question mark indicates that no clear prediction is available in the relevant literature (i.e., the stage is unknown). For the full ATTracTTS-IT dataset, which includes glosses of these examples and additional sentence subtypes, see Appendix A.

ID	Sentence Type	GT	NE	Example (Italian)
i	SV simple	1	1	Alessandro telefona.
ii	SV unaccusative	1	1	Luigi sale.
iii	VS unaccusative	1	1	Arriva Matteo.
iv	Imperatives	1	1	Corri!
v	Modals	1	?	Il babbo vuole saltare.
vi	Root wh-questions	2	1	Chi annaffia i fiori?
vii	Root yes/no questions	2	1	Ha mangiato la mela?
viii	Preposed Adverbs	2	1	Raramente Giorgio dorme.
ix	Focus	2	?	No, l'uccellino salutano i bambini!
x	Illocutionary COMPs	3	1	Che brutto!
xi	Why questions	3	2	Perché il bambino piange?
xii	Topics	3	2	Il cavallo, la bambina lo lava.
xiii	Embedded that	3	2	Il cavallo vede che la mucca beve l'acqua.
xiv	Embedded if	3	2	Non so se Luca verrà al mare.
xv	Subject Relative	3	2	Il bambino che gioca con la mamma.
xvi	Object Relative – intervener	3	2	Il ragazzo che loro abbracciano.
xvii	Object Relative + intervener	4	?	Il ragazzo che la nonna abbraccia.

acquired early. This point is particularly relevant when modeling the developmental trajectory of a language model, whose training, by definition, lacks access to referential stimuli such as here-and-now context (cf. the symbol grounding problem [38]).

3.4. Predictions

Under a NE view, the timing and trajectory of syntactic acquisition are governed by the complexity of formal features involved, rather than its fixed hierarchical position in the functional spine. More specifically, if the GT predicts that Topics (pertaining to Stage 3) are acquired later than wh-questions (pertaining to Stage 2), by virtue of their structural height; from a NE point of view, this depends on the featural specification of these elements [34]: for example, yes/no questions are expected to be early-acquired CP structures due to their low formal complexity and learnability via generalization from minimal cues, whereas according to the GT they are expected to arise in Stage 2.

Under the NE view, the macrocategories C, T, and V are assumed to be available from the onset. In contrast, the GT approach posits that only V and T are initially available (Stage 1), with C-related projections emerging at later stages.

Despite being grounded in empirical studies, the two approaches yield diverging predictions about the order of acquisition. This divergence stems also from how particular structures are analyzed. For instance, whether a given construction involves movement to C or remains within the TP layer is often a matter of theoretical interpretation, and currently under scrutiny.

4. Experimental Evidence

4.1. Methods

To test LMs against the developmental predictions of both NE and GT frameworks, we defined a problem space designed to capture the full range of potential developmental trajectories a LM might exhibit. Using a test set (c.f. next subsection) that targets structurally rich constructions attested at various stages of acquisition, we expect a coherent model (i) to be sensitive to syntactic variations and similarities across different sentence types and to assign probabilities accordingly, and (ii) to align with one of the two developmental hypotheses by assigning higher perplexity scores to items corresponding to later stages of acquisition. To obtain perplexity measures and standard errors, we used the lm-evaluation-harness platform [39] and created a custom task consisting of 100 lexically irrelevant variations of the syntactic patterns presented in Table 1 and further detailed in Appendix A. Items were grouped into three stages to reflect the finer-grained distinctions predicted by the GT framework. If no difference is found between Stage 1 and Stage 2, then the LM behavior is consistent with NE approach. Otherwise, if a distinction emerges, this is in line with GT predictions.

4.2. ATTracTSS: A Novel Dataset

The novel test set we created for evaluating the Acquisition Trajectories of various LMs in Terms of Syntactic Structures is dubbed ATTracTSS. The dataset consists of grammatical sentences representing 9 prototypical syntactic constructions—here referred to as sentence types (e.g., simple SV sentences, wh-questions, topicalizations, embedded clauses)—and 100 lexically diverse items generated for each sentence type.

We built our dataset based on the phenomena tested by GT and NE. Notably, NE does not provide an explicit list of the specific sentence types it predicts to emerge in a fixed acquisitional order. Therefore, we adapted GT's classification to the NE framework where possible, deriving stage-based predictions for both hypotheses Table 1. In cases where alignment was not possible, we assigned the label *unknown*.

4.3. Implementation

We carry out a perplexity analysis starting from the negative log probabilities assigned by the model to each sentence in the dataset. Perplexity levels are expected to inversely correlate with learnability. Perplexity measures how well a model predicts a given sentence. Lower perplexity means the model finds the sentence more predictable (less surprising), while higher perplexity means the model finds it less predictable (more surprising). Given the 100 repetitions of the same syntactic skeleton, we assume that averaging over multiple lexicalizations reduces the impact of individual word-level frequency effects on model perplexity.

At stage level, our hypothesis is that different acquisition stages would be characterized not only by different mean perplexity values, but also by similar standard deviations (SD), indicating consistent model confidence within each stage. As for sentence types, if perplexity remains consistently low across lexical variants of a sentence type, and the variation is low, we interpret this as evidence that the model handles the structure with a degree of robustness and consistency, suggesting it has learned to generalize over that syntactic pattern. While this should not be taken to imply that the model has acquired the structure in a human-like or abstract sense, such behavior can nonetheless serve as a useful proxy for comparison with human acquisition data.

Four models were tested: ita-baseline-small—the pretrained GPT2 baseline model for Italian shared by the BabyLM Community in the HuggingFace platform [10], NeTS-3M—a similar small GPT2 model trained on a custom 3M corpus of child-directed speech [40] —, GePpeTto—117M parameters [11]—and a larger model, Miverva-7B-base, 7B parameters [12]. For the NeTS-3M model we also implemented a longitudinal tracking by repeating the log-probability analysis across multiple training epochs, in order to trace whether the model's familiarization path mirrors human developmental patterns. The same type of analysis could not be carry out on the other models due to the impossibility to carefully control their training.

4.4. Results

Mean perplexity and SD values for each stage in GT and NE were derived from negative log probability values that the four models assigned to each of the items in the dataset, as reported in Table 2 (GT) and Table 3 (NE). Despite numerical differences, perplexity tends to increase coherently with the stage progression in all LMs; SD, instead, tends to grow higher in the latest stages of both GT (Stage 3) and NE (Stage 2), suggesting higher variation within them.

Then, a series of linear regressions were run to assess whether negative log probability assignment is significantly predicted across models (i) by the different syntactic structures of the sentence types included in the dataset, and most importantly (ii) by the articulation in stages proposed by GT and/or by NE. Random intercepts for length (i.e, number of words in each item in the dataset) were included in all regressions. Likelihood ratio tests (ANOVA) between a null model and a model using sentence types as fixed effect revealed that these significantly improved model fit in all LMs (ita-baseline-small: $\chi^2(65) = 2622.7$, $p < .0001$; NeTS-3M: $\chi^2(65) = 2953.7$, $p < .0001$; GePpeTto: $\chi^2(65) = 2925.3$, $p < .0001$; Minerva: $\chi^2(65) = 3095.7$, $p < .0001$). As for GT and NE, instead, similar tests outputted a sharp asymmetry in the predictive power of the two accounts. Treating GT's three-stage articulation as fixed factor significantly improved model fit (ita-baseline-small: $\chi^2(2) = 10.633$, $p < .00491$; NeTS-3M: $\chi^2(2) = 376.68$, $p < .0001$; GePpeTto: $\chi^2(2) = 9.1605$, $p < .0001$; Minerva: $\chi^2(2) = 35.5$, $p < .0001$), but the same did not apply to NE's stages (p values >.05 for all LMs). Note however that except for NeTS-3M, where all pairwise comparisons between stages reach significance, contrasts between Stage 2 and 3 and Stage 1 and 3 strongly vary across LMs (see Appendix B), with Stage 3 being the least stable of the three. For the detailed longitudinal results of the NETS-3M model, see Appendix C.

4.5. Discussion

The experiments reported in the previous sections were conducted to address the issue of language development in LMs, i.e., to assess whether the way LMs "learn" their language may be compared to the process of natural language acquisition in children. Specifically, we compared

Table 2
Mean perplexity estimation and SD grouped by GT stages.

Stages GT	Perplexity (SD)	Models
Overall	42.1788 (13.28)	ita-baseline-small
	50.0302 (16.29)	NeTS-3M
	44.9620 (10.98)	GePpeTto
	36.5133 (11.14)	Minerva
Stage 1	37.3312 (10.28)	ita-baseline-small
	33.7826 (12.35)	NeTS-3M
	40.6229 (8.60)	GePpeTto
	32.3002 (8.62)	Minerva
Stage 2	48.4068 (9.89)	ita-baseline-small
	61.6422 (13.26)	NeTS-3M
	50.5393 (8.33)	GePpeTto
	41.3775 (8.41)	Minerva
Stage 3	55.2353 (17.35)	ita-baseline-small
	65.0507 (17.23)	NeTS-3M
	56.5017 (12.70)	GePpeTto
	48.5069 (13.62)	Minerva

Table 3
Mean perplexity estimation and SD grouped by NE stages.

Stages NE	Perplexity (SD)	Models
Overall	42.1788 (13.28)	ita-baseline-small
	50.0301 (16.29)	NeTS-3M
	44.9620 (10.98)	GePpeTto
	36.5133 (11.14)	Minerva
Stage 1	38.8547 (10.84)	ita-baseline-small
	46.4309 (14.82)	NeTS-3M
	41.8718 (8.85)	GePpeTto
	33.4046 (8.45)	Minerva
Stage 2	54.8328 (15.47)	ita-baseline-small
	66.5525 (16.33)	NeTS-3M
	57.1496 (13.13)	GePpeTto
	48.0398 (14.31)	Minerva

the stage-wise developmental predictions of two competing theories, GT and NE, against the performance of some Italian LMs. We did that by looking at perplexity associated to a varied set of sentences in a novel dataset (ATTracTSS test set) both in a cross-sectional perspective, looking at four different Italian models (ita-baseline-small, NeTS-3M, GePpeTto, Minerva), and in a longitudinal perspective, focusing on the performance of one of these models (NeTS-3M) across training epochs.

As for the cross-sectional study, we observed a general alignment of all our LMs with the linguistic development observed in children. Perplexity values tended to grow with the progression of stages in both GT and NE, suggesting that the syntactic structures that children struggle with the most—and therefore take longer to be acquired—roughly overlap with the sentence types that LMs find less predictable. Nevertheless, closer inspection of mean perplexity values per sentence type revealed some variation within the stages, especially in GT 3 and NE 2: some late structures for children, like why-questions, receive very low perplexity from all models (~30), while Stage 1 transitive clauses are assigned higher-than expected perplexity (~52). These observation suggest that caution is needed when comparing humans and LMs, and while the general learning trend aligns with human acquisition, some important asymmetries remain.

Moreover, and as a general consideration, our results show consistently higher perplexity if compared to standard benchmarks (e.g., ~20 perplexity for GPT-3 [41]). This may stem from the absence of licensing contexts in the test items, or suggest that the models resolve, for instance, certain non-local dependencies—especially those in the Left Periphery—via strategies that diverge from native-like structural processing. Also, this suggests that our assessment task is far from trivial and highlights the need for further exploration of training regimens to determine whether specific language models exhibit learning trajectories consistent with those observed in human language acquisition.

Another interesting result concerns the difference in the predictive power of GT and NE with respect to LMs performance. While the single structure types always qualify as good predictors for LMs perplexity, ratifying some sort of syntactic representation abilities, grouping the phenomena into the three-stage articulation of GT always returns better results than the coarser two-stage subdivision of NE. This pattern holds both across models, and in the longitudinal evaluation across training epochs of NeTS-3M, a small-scale transformer trained on 3M tokens of child-directed speech. Even though proposed on independent grounds, then, the linguistic stages grounded in the GT framework may offer a useful lens for interpreting LM behavior, especially in cognitively oriented settings.

Finally, two more relevant considerations may be drawn especially from the epoch-by-epoch analysis, which we could performed on NeTS-3M, the only model we could strictly control for architecture, training regimen and training set (a 3M token corpus, including child-directed speech only, [40], see Appendix C).

First, the model shows evidence of learning, gradually reducing the perplexity gap between items from GT Stage 2 and Stage 3, although these stages remain distinguishable. However, and in line with the results of pairwise comparisons between Stages across models, the most pronounced distinction for the model clearly lies between Stage 1 and Stage 2.

Second, our findings suggest that minimal (3M tokens of NeTS-3M vs. $\geq$ 10M tokens of the other LMs) but curated input allows a transformer model to approximate early, mid, and late stages of language acquisition, in line with empirically attested developmental pattern of a linguistic theory (the GT approach): this is confirmed not only by general the snapshot of perplexity estimation across stages, where NeTS-3M is the only LM strongly differentiating Stage 1, 2 and 3, but crucially also along training epochs simulating child linguistic development.

5. Concluding remarks

In this paper we presented ATTracTSS, a novel dataset to assess the Acquisition Trajectories in Terms of Syntactic Structures inspired by language acquisition studies and by two competing empirically-grounded theories—the Growing Trees approach and the Neo-Emergentist framework. Both theories argue for a stage-wise acquisition of syntax in children, but crucially differ in the size and internal composition of these stages.

We conducted out-of-the-box evaluations on three "small" language models (124M parameters)—the pretrained GPT2 baseline model for Italian shared by the BabyLM, ita-baseline-small; NeTS-3M, a similar small GPT2 model trained on a custom 3M corpus; GePpeTto, 117M parameters—and a larger model, Miverva-7B-base, 7B parameters. We measured the perplexity that these LMs assigned to each sentence in the test set and compared them against the three-stage predictions of GT and the two-stage articulation of NE.

In our experimental results, we observe that small-scale, fully open Italian-tuned models show alignment with theories of language acquisition. Among the theoretical approaches tested, the GT-based stage theory yields more accurate predictions than the NE-based approach. This work demonstrates the benefits of a sufficiently rich grammatical theory in order to account for how language acquisition unravels in children, and how this developmental trajectory can serve as a metric to compare natural, instinct-driven acquisition in humans with the learning processes of LMs. In children, acquisition proceeds incrementally, through identifiable phases or stages. A robust theory is necessary to explicitly determine which linguistic phenomena emerge at which stage, in a principled and non-impressionistic way. Such reflection is crucial for linguists, but it may also have broader practical implications, particularly with respect to the sustainability and optimization of model training. Focusing on child language acquisition, and especially on the stages through which it unfolds, offers an additional, more fine-grained metric for evaluating model competence and cognitive plausibility. In this work, we did not address the notion of cognitive coherence, which we consider too general; instead, we focus on strictly linguistic issues and discuss structural coherence with respect to native speaker intuitions—where structural refers specifically to syntactic structures, i.e., sentence types.

This ultimately frames the core tension in terms of *learning* versus *acquisition*.

6. Limitations

Although our aim was to assess acquisition stages also across different training regimens—naturalistic, conversational, or redundant [5] using small-scale corpora (10–100M tokens), we could only test the NeTS-3M model under the redundant regimen, trained on a 3M-token corpus of Italian child-directed speech [40]. This is below the 10M-token BabyLM small track threshold [10]. While minimal, this amount approximates the linguistic input received by a 4-year-old child—who has typically acquired structures across all three developmental stages [42]—though an additional million tokens would have brought the exposure closer to that developmental window.

Moreover, the model architecture—GPT-2 (see model card [43])—is not cognitively plausible, as it relies on a non-incremental, parallel attention mechanism that does not reflect human-like structure-building [2, 5].

A further limitation is that the dataset is not fully balanced in terms of the number of phenomena per item; future iterations will aim to expand the dataset and ensure more uniform distribution across phenomena (see the material in the Appendix A).

Finally, although we draw on attested acquisition patterns from Growing Trees and Neo-Emergentism, we lack adult acceptability data for the same structures. Such data will be essential in future studies to assess whether model outputs at later training stages simulate adult linguistic competence.

Acknowledgments

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title T-GRA2L: Testing GRAdeness and GRAmmaticality in Linguistics – CUP I53D23003900006 - Grant Assignment Decree No. 104 adopted on the 2nd February 2022 by the Italian Ministry of Ministry of University and Research (MUR). PI: Cristiano Chesi.

References

[1] A. Srivastava, A. Rastogi, A. Rao, A. A. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, Transactions on machine learning research (2023).

[2] C. Chesi, M. Barbini, V. Bressan, S. Neri, M. L. Piccini Bianchessi, S. Rossi, T. Sgrizzi, Different Ways to Forget: Linguistic Gates in Recurrent Neural Networks, in: M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt, E. G. Wilcox (Eds.), Proceedings of the BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, 2024. URL: https://aclanthology.org/2024.conll-babylm.9/.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, Virtual Event Canada, 2021, pp. 610–623. URL: https://dl.acm.org/doi/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[4] C. Chesi, A conclusive remark on linguistic theorizing and language modeling, 2025. URL: https://arxiv.org/abs/2506.03268. doi:10.48550/ARXIV.2506.03268, version Number: 1.

[5] C. Chesi, M. Barbini, V. Bressan, A. Fusco, S. Neri, M. L. Piccini Bianchessi, S. Rossi, T. Sgrizzi, From Recursion to Incrementality: Return to Recurrent Neural Networks, Linguistic Vanguard (forthcoming).

[6] N. Friedmann, A. Belletti, L. Rizzi, Growing trees: The acquisition of the left periphery, Glossa: a journal of general linguistics 6 (2021). URL: https://www.glossa-journal.org/article/id/5877/. doi:10.16995/glossa.5877, number: 1.

[7] N. Bosch, Not all complementisers are late: A first look at the acquisition of illocutionary complementisers in Catalan and Spanish, Isogloss. Open Journal of Romance Linguistics 9 (2023) 1–39. URL: https://revistes.uab.cat/isogloss/article/view/v9-n1-bosch. doi:10.5565/rev/isogloss.313, number: 1.

[8] A. Belletti, M. T. Guasti, The Acquisition of Italian: Morphosyntax and its interfaces in different modes of acquisition, volume 57, John Benjamins Publishing Company, Amsterdam, 2015.

[9] W. De Vries, M. Nissim, As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 836–846. URL: https://aclanthology.org/2021.findings-acl.74. doi:10.18653/v1/2021.findings-acl.74.

[10] L. Charpentier, L. Choshen, R. Cotterell, M. O. Gul, M. Hu, J. Jumelet, T. Linzen, J. Liu, A. Mueller, C. Ross, R. S. Shah, A. Warstadt, E. Wilcox, A. Williams, BabyLM Turns 3: Call for papers for the 2025 BabyLM workshop, 2025. URL: http://arxiv.org/abs/2502.10645. doi:10.48550/arXiv.2502.10645, issue: arXiv:2502.10645 arXiv:2502.10645 [cs].

[11] L. De Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, GePpeTto Carves Italian into a Language Model, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, CEUR-WS.org, Bologna, 2021.

[12] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR, Aachen, 2024.

[13] M. T. Guasti, Language acquisition: the growth of grammar, second edition ed., The MIT Press, Cambridge, MA, 2016.

[14] S. Crain, R. Thornton, Investigations in universal grammar: a guide to experiments on the acquisition of syntax and semantics, Language, speech and communication, MIT, Cambridge, Mass., 2000.

[15] N. Chomsky, Barriers, MIT Press, Cambridge, MA, 1986.

[16] S. T. Piantadosi, F. Hill, Meaning without reference in large language models, 2022. URL: https://arxiv.org/abs/2208.02957. doi:10.48550/ARXIV.2208.02957, version Number: 2.

[17] S. T. Piantadosi, Modern language models refute Chomsky's approach to language, in: E. Gibson, M. Poliak (Eds.), From fieldwork to linguistic theory: A tribute to Dan Everett, Language Science Press, Berlin, 2024. URL: https://zenodo.org/doi/10.5281/zenodo.12665933. doi:10.5281/ZENODO.12665933.

[18] A. Goyal, Y. Bengio, Inductive biases for deep learning of higher-level cognition, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 478 (2022) 20210068. URL: https://royalsocietypublishing.org/doi/10.1098/rspa.2021.0068. doi:10.1098/rspa.2021.0068, number: 2266.

[19] J. Fodor, The Modularity of Mind., The Philosophical Review 94 (1985) 101. URL: https://www.jstor.org/stable/2184717?origin=crossref. doi:10.2307/2184717, number: 1.

[20] G. Baggio, A. De Santo, N. A. Nuñez, Plausibility and Early Theory in Linguistics and Cognitive Science, Computational Brain & Behavior 7 (2024) 535–547. URL: https://link.springer.com/10.1007/s42113-024-00196-7. doi:10.1007/s42113-024-00196-7, number: 4.

[21] L. Rizzi, Grammatically-Based Target-Inconsistencies in Child Language, in: K. U. Deen, J. Nomura, B. Schulz, B. D. Schwartz (Eds.), The Proceedings of the Inaugural Conference on Generative Approaches to Language Acquisition—North America, MIT Working Papers in Linguistics, 2006.

[22] L. Rizzi, Some Notes on Linguistic Theory and Language Development: The Case of Root Infinitives, Language Acquisition 3 (1993) 371–393. URL: http://www.tandfonline.com/doi/abs/10.1207/s15327817la0304_2. doi:10.1207/s15327817la0304_2, number: 4.

[23] L. Haegeman, Root Infinitives, Tense, and Truncated Structures in Dutch, Language Acquisition 4 (1995) 205–255. URL: http://www.tandfonline.com/doi/abs/10.1207/s15327817la0403_2. doi:10.1207/s15327817la0403_2, number: 3.

[24] M. Salustri, N. Hyams, Looking for the universal core of the RI stage, in: V. Torrens, L. Escobar (Eds.), Language Acquisition and Language Disorders, volume 41, John Benjamins Publishing Company, Amsterdam, 2006, pp. 159–182. URL: https://benjamins.com/catalog/lald.41.09sal. doi:10.1075/lald.41.09sal.

[25] L. Rizzi, G. Bocci, Left Periphery of the Clause: Primarily Illustrated for Italian, in: M. Everaert, H. C. Riemsdijk (Eds.), The Wiley Blackwell Companion to Syntax, Second Edition, 1 ed., Wiley, 2017, pp. 1–30. URL: https://onlinelibrary.wiley.com/doi/10.1002/9781118358733.wbsyncom104. doi:10.1002/9781118358733.wbsyncom104.

[26] A. Radford, Syntactic theory and the acquisition of English syntax: The nature of early child grammars of English. Blackwell: Oxford., Blackwell, Oxford, 1990.

[27] A. Belletti, N. Friedmann, L. Rizzi, Growing trees in child grammars: Cartography as an analytic tool for syntactic development, in: S. Wolfe (Ed.), The Oxford Handbook of Syntactic Cartography, ????

[28] G. Cinque, L. Rizzi, The cartography of syntactic structures, in: B. Heine, H. Narrog (Eds.), The oxford handbook of linguistic analysis, Oxford University Press, Oxford, 2010, pp. 65–78.

[29] S. Rossi, Italian/Romance imperatives as radically reduced structures: a corpus CHILDES study, RGG 45 (2023) 1–39. Number: 5.

[30] E. Casadei, A New Sentence Repetition Task Tool to Investigate The Acquisition of Syntactic Structures in Typical and Atypical Development: A View From Growing Trees and Syntactic Cartography, Master's thesis, University of Siena, Siena, 2024.

[31] T. Sgrizzi, When infinitives are not under control: the Growing Trees Hypothesis and the developmental advantage of restructuring verbs, RGG 46 (2024) 1–39. Number: 4.

[32] A. A. Robiatu, A Computational Perspective on The Growing Tree Approach: Design and Implementation of A Rule-Based System, Master's thesis, University of Siena, 2025.

[33] N. Bosch, T. Biberauer, Emergent Syntactic Categories and Increasing Granularity: Evidence from a Multilingual Corpus Study, in: Proceedings of the 48th Boston University Conference on Language Development (BUCLD), Cascadilla Proceedings Project, Somerville, MA, 2024, pp. 101–116.

[34] N. Bosch, Not all topics are equal: syntactic complexity and its effect on the acquisition of left-peripheral structures, in: Proceedings of NELS 55, 2024.

[35] N. Chomsky, Three Factors in Language Design, Linguistic Inquiry 36 (2005) 1–22. URL: https://direct.mit.edu/ling/article/36/1/1-22/250. doi:10.1162/0024389052993655, number: 1.

[36] L. Rizzi, Early null subjects and root null subjects. in Syntactic theory and first language acquisition: Cross-linguistic perspectives„ in: Binding, dependencies, and learnability., Lawrence Erlbaum Associates Inc., Hillsdale, NJ, 1994.

[37] J. Heim, M. Wiltschko, Rethinking structural growth: Insights from the acquisition of interactional language, Glossa: a journal of general linguistics 10 (2025). URL: https://www.glossa-journal.org/article/id/16396/. doi:10.16995/glossa.16396, number: 1.

[38] J. R. Searle, Minds, brains, and programs, Behavioral and Brain Sciences 3 (1980) 417–424. URL: https://www.cambridge.org/core/product/identifier/S0140525X00005756/type/journal_article. doi:10.1017/S0140525X00005756, number: 3.

[39] L. Sutawika, H. Schoelkopf, L. Gao, B. Abbasi, S. Biderman, J. Tow, B. Fattori, C. Lovering, et al., Eleutherai/lm-evaluation-harness: v0.4.9.1, 2025. doi:10.5281/ZENODO.16737642.

[40] A. Fusco, M. Barbini, M. L. Piccini Bianchessi, V. Bressan, S. Neri, S. Rossi, T. Sgrizzi, C. Chesi, Recurrent Networks are (Linguistically) Better? An Experiment on Small-LM Training on Child-Directed Speech in Italian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR, Aachen, 2024.

[41] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam,

G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, 2020. URL: http://arxiv.org/abs/2005.14165. doi:10.48550/arXiv.2005.14165, issue: arXiv:2005.14165 arXiv:2005.14165 [cs].

[42] N. Friedmann, J. Reznick, Stages rather than ages in the acquisition of movement structures: Data from sentence repetition and 27696 spontaneous clauses, Glossa: a journal of general linguistics 39 (2021). URL: https://www.glossa-journal.org/article/id/5716/. doi:10.16995/glossa.5716, number: 1.

[43] M. Ö. Gül, babylm-baseline-10m-gpt2, https://huggingface.co/BabyLM-community/babylm-baseline-10m-gpt2, 2025. Model card last updated ca. 1 month before August 2025.

A. Online Resources

Additional resources, including the full ATTracTSS dataset and supporting materials, are available at:

- ATTracTTS GitHub repository

To stay up to date with future developments from our lab, visit:

- NeTS Lab - Computational Projects
- NeTS Lab - General website

B. Pairwise Comparisons

This appendix reports the results of pairwise statistical comparisons between the estimated probabilities associated with each GT stage. See Table 4.

C. NeTS-3M Model Results Across Epochs

This appendix reports the results of the NeTS-3M Model across epochs, see Table 5. We show performance over 10 training epochs, with predictions evaluated by linguistic phenomenon, GT approach, and NE approach. Table 5 reports perplexity (derived from -log(probability), where higher values indicate greater model uncertainty) and χ^2 values (where higher values reflect stronger model predictions).

Table 4
Pairwise comparisons between estimated probabilities of GT Stages.

Models	Contrast	Est	SE	*p* value
ita-baseline-small	Stage 1 vs. Stage 2	0.818	0.270	.007
	Stage 2 vs. Stage 3	-0.048	0.367	.991
	Stage 1 vs. Stage 3	-0.866	0.377	.056
NeTS-3M	Stage 1 vs. Stage 2	6.710	0.341	$<.001$
	Stage 2 vs. Stage 3	2.800	0.464	$<.001$
	Stage 1 vs. Stage 3	3.910	0.476	$<.001$
GePpeTto	Stage 1 vs. Stage 2	0.481	0.189	.029
	Stage 2 vs. Stage 3	-0.180	0.257	.762
	Stage 1 vs. Stage 3	-0.661	0.263	.032
Minerva	Stage 1 vs. Stage 2	-0.934	0.188	$<.001$
	Stage 2 vs. Stage 3	-1.215	0.255	$<.001$
	Stage 1 vs. Stage 3	-0.281	0.262	.532

Table 5
Performance of the NeTS-3M model over 10 training epochs. Deeper green indicates better performance.

	Average Perplexity: -log(prob)				**Linear Mixed Effects fitted Models**					
					By phenomenon		**By GT predictions**		**By NE predictions**	
Epoch	**Total**	**Stage 1**	**Stage 2**	**Stage 3**	χ^2**(65)**	**p**	χ^2**(3)**	**p(GT)**	χ^2**(2)**	**p(NE)**
1	57,11293	46,85316	72,95726	76,02506	2740,6	<.00001	360,37	<.00001	11,608	0,00006
2	55,92298	46,2499	70,86065	73,75493	2753,3	<.00001	395,08	<.00001	19,37	0,00001
3	53,18674	43,55008	68,48524	70,06534	3008,9	<.00001	535,53	<.00001	7,5447	0,006
4	50,67562	41,39614	65,38968	66,96554	3016,6	<.00001	563,55	<.00001	10,459	0,001
5	50,61927	41,65151	64,78868	66,46903	2964,9	<.00001	485,05	<.00001	5,9709	0,0145
6	50,07034	41,41508	63,71472	65,43424	3078,4	<.00001	501,18	<.00001	7,0212	0,008
7	50,50437	42,14307	63,68911	65,3385	2927,2	<.00001	394,24	<.00001	4,2885	0,03837
8	49,87423	41,54157	63,21636	64,22711	3009,7	<.00001	438,69	<.00001	0,8596	0,3538
9	48,38463	40,33959	61,03486	62,7337	2867,9	<.00001	339,63	<.00001	4,4175	0,03557
10	48,56504	40,68987	61,11143	62,2642	2953,7	<.00001	376,68	<.00001	3,0096	0,08