What Kind of Language is Easy to Language-Model Under Curriculum Learning?

Nadine El-Naggar, Tatsuki Kuribayashi, Ted Briscoe
Mohamed bin Zayed University of Artificial Intelligence
{nadine.naggar, tatsuki.kuribayashi, ted.briscoe}@mbzuai.ac.ae

Abstract

Many of the thousands of attested languages share common configurations of features, creating a spectrum from typologically very rare (e.g., object-verb-subject word order) or impossible languages to very common combinations of features (e.g., subject-object-verb word order). One central question is under what conditions such typological tendencies can be predicted, and specifically whether the learning bias of language models (LMs) is sufficient to reproduce such patterns. In this study, we add one dimensionality to such analysis — the learning scenario for LMs — to explore its interaction with the inductive bias of LMs. Specifically, as a first study, we examine the effect of curriculum learning (CL), as a developmentally motivated learning scenario, i.e., starting with simpler sentences rather than randomly-ordered input. We expand existing LM-based exploration (El-Naggar et al., 2025a,b) with a simple CL variant and find that CL substantially impacts the apparent inductive bias of LMs.

Keywords: Artificial languages, curriculum learning, generalization

1. Introduction

Natural languages (NLs) exhibit a range of properties, including different word order configurations, raising the question: which, if any, types of languages are easier for language models (LMs) to learn (Cotterell et al., 2018; Mielke et al., 2019; White and Cotterell, 2021; Borenstein et al., 2024; Arnett and Bergen, 2025)? Related questions include why some grammatical feature combinations are typologically common while others are rare (Dryer and Haspelmath, 2013), and to what extent LMs' domain-general learning biases replicate these linguistic tendencies (Chomsky et al., 2023; Xu et al., 2025). Existing studies have reported that, for example, the inductive bias of a specific class of LMs aligns with typological commonality, and the properties of such typologically-aligned models (e.g., working memory limitations) could partially explain such typological tendencies (Kuribayashi et al., 2024; El-Naggar et al., 2025a,b). Although such computational simulations of language learnability can serve as proof-of-concept for the idea that learning biases shaping language (Culbertson et al., 2012), LMs' and humans' language acquisition settings are generally different (e.g., the amount of data), and to make such simulations more relevant to human language science, it is necessary to align LM training scenarios with humans' ones (Warstadt and Bowman, 2022).

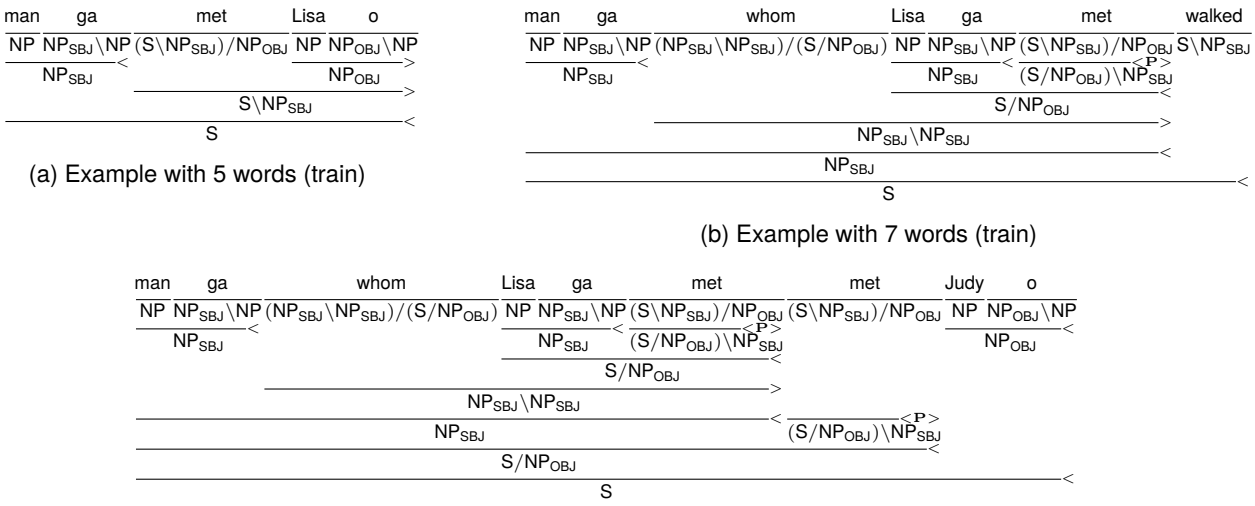
One factor that is relevant to human language acquisition and overlooked in existing LM-based simulations of typological patterns is the order of data presentation. For example, children with severe working memory limitations may take in relatively simple sentences first through language acquisition (Hudson Kam and Newport, 2005; Kam and Newport, 2009). Such effects have typically been simulated with curriculum learning (CL) in computational simulations, which present training data from simpler to more complex sentences.

In this study, we explore one basic CL setting in the LM-based simulation of typological patterns. That is, our research question is *what kind of language is easier for LMs to learn under CL*. As an initial foray, we adopt a simple length-based CL (Spitkovsky et al., 2009) and replicate existing studies (El-Naggar et al., 2025a,b) to analyze interaction effect between CL and LMs' inductive biases over diverse word orders. Our experimental results demonstrate that the word-order preferences indeed change with the CL setting, and, somewhat surprisingly, the CL-based results deviate more from typological commonality, which raises several implications (§ 5).

2. Background

Artificial Language Learning Artificial languages (ALs) are often used in the evaluation of LMs as they allow for more targeted evaluation of specific linguistic features. Typical investigations include analyzing what kind of structure LMs can model, using ALs of varying complexities (Suzgun et al., 2019; Weiss et al., 2018; El-Naggar et al., 2022; Kallini et al., 2024; Someya et al., 2024; Delétang et al., 2023). In addition to formal languages (e.g., Dyck languages), more linguistically grounded ALs, such as PCFG-based ones, have been developed to evaluate LM word-order inductive biases (White and Cotterell, 2021). El-Naggar et al. (2025a,b) extended PCFG-based ALs to generalized categorial grammar (GCG)-based

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 48–59
16 May 2026.



(a) Example with 5 words (train)

(b) Example with 7 words (train)

(c) Example with 9 words (MEDIUM test)

Figure 1: Examples of sentences and their GCG derivation (somewhat simplified for space limitations).

ALs, which support a broader range of grammatical phenomena. ALs are also often used to test LM generalization under controlled conditions (Weiss et al., 2018; Suzgun et al., 2019; El-Naggar et al., 2022, 2023, 2025b).

Curriculum Learning Research has long questioned whether the order in which training data is exposed to neural network models affects their learning (Elman, 1991, 1993; Rohde and Plaut, 1999; Krueger and Dayan, 2009; Bengio et al., 2009). Elman (1991, 1993) introduced the idea of "starting small" and conducted experiments where RNNs were trained in phases to learn English sentences, where sentence complexity was increased in each phase. The starting small hypothesis has been revisited in several NLP applications, like learning and generalization of grammatical patterns (Rohde and Plaut, 1999), unsupervised dependency parsing (Spitkovsky et al., 2009) and in a reinforcement learning framework (Krueger and Dayan, 2009), and even in visual applications like shape recognition (Bengio et al., 2009).

3. Experimental Settings

3.1. Original Setting

We first briefly introduce the base datasets adapted from prior works (El-Naggar et al., 2025a,b). A set of GCG grammars is first defined via multiple independent word-order parameters generating 96 languages for the GCG-based AL corpus. For example, one AL follows Japanese-like (head-final), and another follows English-like (mostly head-initial) word order. Specifically, there are 7 binary parameters to determine (i) subject–verb order, (ii) object–verb order, (iii) subject–object order, (iv) complementizer–clause, (v) noun-adposition, (vi) noun-adjective, and (vii) position of relativizer. These configurations are denoted as a sequence of digits (e.g., `0001010`). Configurations with more `0`s tend to be head-final, while the ones with more `1`s are head-initial. The binary parameter controls the directional slash of specific rules in the GCG grammar. Examples of the sentence "Kim said that John touched Lisa" in grammars `0000000` (Japanese-like) and `0101101` (English-like) are:[1]

`0000000`: *Ken ga John ga Lisa o touch that said*

`0101101`: *Ken ga said that John ga touch Lisa o*

Note that the ALs have case markers for subject (*ga*) and object (*o*). Figure 1 shows more examples of sentences derived by `0101101` (English-like) GCG grammar. Note that sentences are first generated fully randomly, and then grammatical (parsable) ones are selected based on the GCG parser, where GCG parser configurations differ in different word order configurations.

In each AL, the training data consists of 80K sentences with 3–8 words (uniform length distribution), and the LMs' inductive biases are evaluated on three types of evaluation sets: (i) SHORT test set with the same length distribution as the training data, (ii) MEDIUM test set with longer length distribution than the training set (9-10) to test generalization, and (iii) LONG test set that has further longer sentences created by several heuristics (11-20). [2] We follow this data split, but change how

[1] Note that while the words in the ALs are pseudo-words, real English words are used in these examples for readability.

[2] These are available at https://github.com/nadineelnaggar/gcg-based-artificial-languages

the training data is presented to the model and observe its effect on word-order preferences. We define ORIGINAL setting as the model training using this base dataset with a randomized data order.

3.2. Curriculum Learning Setting

To examine the CL effect on LMs' inductive biases, we introduce the CL setting in which the training data is presented based on a specific strategy. As a case study, we explore the widely-used token length-based CL, and thus our questions will be: do LMs prefer specific word order configurations independently of the ordering of training data? Or more generally, how reproducible are existing findings under different LM training scenarios? We do not aim to find the best, optimal learning scenario to control LMs' learning preferences or to perform a comprehensive analysis, trying every possible learning strategy.

Specifically, we split the base training data into six parts (lengths of 3, 4, 5, 6, 7, and 8) with the same number of sentences in each part. LMs are trained sequentially in stages corresponding to each training data part. That is, the first training stage is with sentences of length 3, and the second stage will be with sentences of length 4, and so on. To avoid catastrophic forgetting of previous data, we sample 5% of the data from each past part and inject it into the current part. For example, the data in the third stage consists of length-3 (5%), length-4 (5%), and length-5 (90%) sentences. For a fair comparison with the ORIGINAL setting, the total number of training sentences is set to match the base dataset (80K). For each of 96 ALs in the base dataset, we created a CL trained model.

4. Experiments

We adopt three evaluation settings and replicate El-Naggar et al. (2025b). For non-CL, ORIGINAL settings, results are excerpted from that study for a comparison, and our new CL-based results are superimposed on them. We focus on RNN, LSTM, and Transformer LMs.

Model Training. For the CL settings, we adopt the same setting as the original ones, except for CL (see Appendix C for hyperparameter details). We train LMs with three different random seeds for each model, using the Fairseq toolkit (Ott et al., 2019). Under CL, LMs are trained for 2 epochs in each stage, except the last, and the optimizer (e.g., learning rate scheduler) is inherited across all the stages. At the last stage, an early-stopping criterion of 5-epoch patience is adopted (i.e., if the validation PPL is not improved for 5 epochs, the training stops), following the non-CL settings.

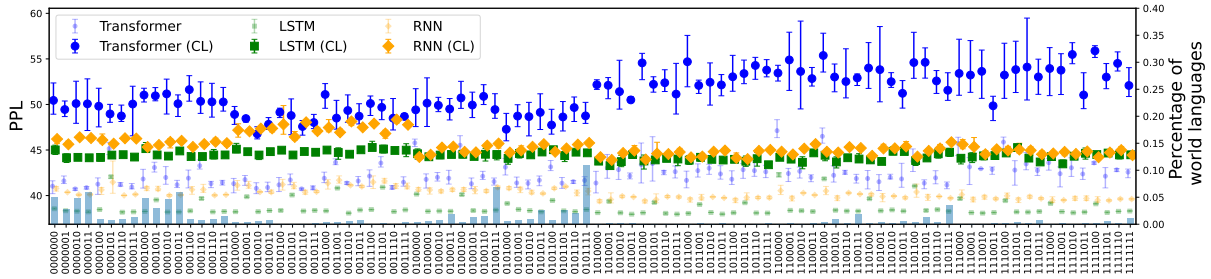
Evaluation. The SHORT test set evaluates the LMs' in-distribution modeling ability, and MEDIUM and LONG ones evaluate the out-of-distribution generalization ability of LMs. Note that the sentences in the SHORT test sets are unseen in the training data. We report average perplexity (PPL) among three runs for each word order configuration. We also report typological alignment (TA) scores — Pearson's correlation coefficient between model's average PPL and the frequency of respective word order configurations in the world (Kuribayashi et al., 2024; El-Naggar et al., 2025a,b). A negative TA score reflects a better alignment of a LM's learning preference with typological tendencies, i.e., typologically common word orders are easier to learn for the LMs. Our focus is on how the language-learning curriculum (ORIGINAL vs. CL) affects their learning preferences (PPLs and TAs).

4.1. Experiment 1: PPLs

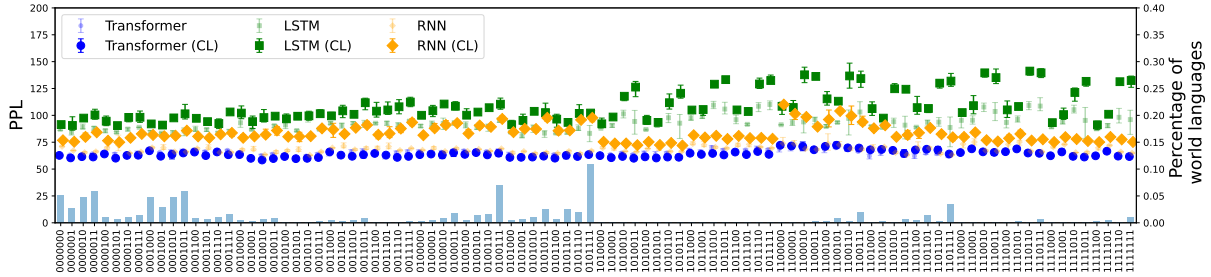
PPL variations and CL. We first evaluate LMs based on perplexity (PPL) on the SHORT, MEDIUM, and LONG test sets. Figure 2 shows the general PPL tendencies (y-axis) over different word order configurations (x-axis) in each test set. Here, PPLs from LMs without CL are also shown with smaller, semi-transparent markers.[3] As a quick check, we first confirm that word order preference still varies even when data are presented in a comprehensive way with CL. Compared to the original results (smaller, semi-transparent markers in Figure 2), we observe that LMs with CL tend to exhibit worse PPL on SHORT tests, but comparable or better PPL on MEDIUM and LONG tests, suggesting less overfitting to short sentences and a positive effect of CL for length generalization.

Typological variations and CL. Table 1 aggregates the results of our CL-based LMs, and the TA score is also computed. To summarize the results, first, the TA scores, especially in RNN/LSTM results on MEDIUM/SHORT sets, substantially change with and without CL, and this basically leads to a worse TA. Second, the TA scores in LONG set are still significantly negative; that is, typologically common word order configurations facilitate generalization to longer sentences, regardless of CL. Therefore the learning setting does affect the results, and in our case, this impact was not equally strong nor random across models/conditions, but rather altered specific results on individual ALs and test sets. Further experimentation with different approaches to CL, such as ordering by construction

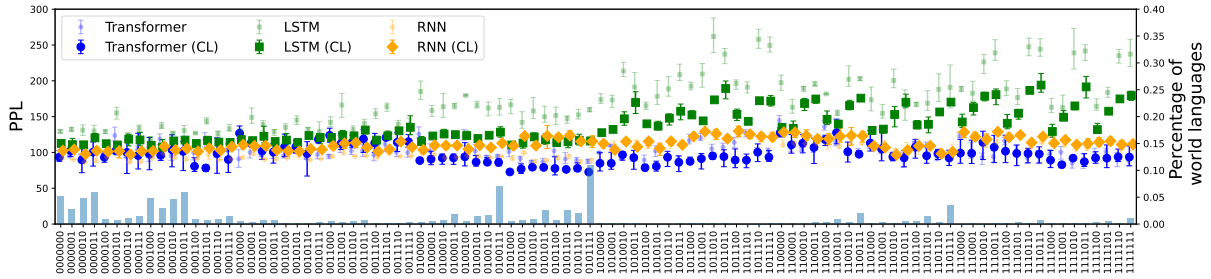
[3] These scores without CL are exactly the same as those reported in El-Naggar et al. (2025b). Our CL results are comparable with their study as the adaptation of CL is the only difference.



(a) SHORT test (length 3–8).



(b) MEDIUM test (length 9–10).



(c) LONG test (length 11–20).

Figure 2: Distributions of perplexities and typological plausibility across languages. The error bars indicate max and min PPLs within three runs. The smaller, semi-transparent markers correspond to the results without CL, which is excerpted from El-Naggar et al. (2025b).

complexity as opposed to token length, may shed further light on the interactions between models, learning scenarios and inductive bias.

Inter-model correlations. Figure 3 shows inter-correlation of PPL-vectors over 96 word orders from different LMs and CL settings. Note that the orthogonal elements in the matrices show the average correlation between PPL-vectors from the same LM architecture, but with different seeds. First, PPL correlation between the same model with and without CL is generally smaller than the seed variance (respective orthogonal elements), showing that CL does affect the word order preference of LMs. Second, CL has different effects depending on LM architectures; for example, in SHORT test, LSTM vs. LSTM (CL) shows nearly zero correlation (0.036), in MEDIUM test, RNN vs. RNN (CL) shows a small correlation (0.145), while in LONG test, Transformer vs. Transformer (CL) shows a relatively small correlation (0.276). This demonstrates subtle and non-obvious interactions between word order preferences, model architectures, and learning scenarios.

4.2. Experiment 2: Targeted Evaluations

We also replicate the targeted evaluations conducted in El-Naggar et al. (2025b) (see details in their Sections 6 and 7).

4.2.1. PPLs in Targeted Generalization Sets.

We measure PPLs on the test data with specific complex constructions. Here, the generalization is evaluated on unbounded dependency structures, specifically recursive relative clauses (RECURSIVE), where relative clauses are nested, and embedded relative clauses (EMBEDDED), where a relative clause is embedded in a subordinate clause. English examples are as follows:

Recursive Relative Clauses: *fruits ga which pasta ga which John ga promised nibbles received wall o*

Model	CL	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓
		Short						
Trans. (El-Naggar et al., 2025b)		41.8	41.6	42.3	42.6	42.7	43.3	**−27.7**†
Trans.	✓	50.2	48.8	49.3	52.6	53.3	53.4	**−22.3**†
LSTM (El-Naggar et al., 2025b)		38.7	38.8	38.7	38.4	39.1	38.5	**−14.2**
LSTM	✓	44.4	44.9	44.5	43.9	44.2	44.4	16.1
RNN (El-Naggar et al., 2025b)		40.4	41.0	40.6	39.7	40.1	39.7	13.0
RNN	✓	45.9	47.4	45.1	44.5	45.0	44.8	14.2
NL (Prob. ↑)		0.54	0.04	0.23	0.01	0.12	0.05	-

(a) Short test set.

Model	CL	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓
		Medium						
Trans. (El-Naggar et al., 2025b)		65.2	63.5	64.2	65.9	66.1	65.0	**−10.4**
Trans.	✓	63.2	61.8	62.9	62.8	68.9	64.7	**−5.9**
LSTM (El-Naggar et al., 2025b)		85.9	91.7	88.0	97.5	92.9	97.9	**−31.0**†
LSTM	✓	95.6	102.3	101.9	112.4	119.8	117.7	**−20.1**†
RNN (El-Naggar et al., 2025b)		67.8	67.9	66.7	69.6	69.0	69.4	**−17.4**
RNN	✓	80.3	84.5	89.7	76.6	91.7	78.1	21.2
NL (Prob. ↑)		0.54	0.04	0.23	0.01	0.12	0.05	-

(b) Medium test set.

Model	CL	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓
		Long						
Trans. (El-Naggar et al., 2025b)		102.3	99.4	97.9	104.0	107.6	97.9	**−19.2**
Trans.	✓	95.1	112.9	83.1	89.9	106.2	96.2	**−18.0**†
LSTM (El-Naggar et al., 2025b)		131.9	141.5	160.7	205.5	180.9	207.5	**−33.4**†
LSTM	✓	113.8	122.0	119.3	153.6	151.9	163.7	**−32.3**†
RNN (El-Naggar et al., 2025b)		91.8	94.6	93.2	118.0	109.0	114.2	**−43.1**†
RNN	✓	102.9	107.2	113.0	117.6	113.7	117.8	**−20.2**†
NL (Prob. ↑)		0.54	0.04	0.23	0.01	0.12	0.05	-

(c) Long test set.

Table 1: Average PPLs within each base word order group and Pearson's correlation coefficient (TA) between PPL and typological frequency. Negative TA scores are highlighted in bold. Statistical significance ($p<0.05$) is marked with †. NL is the percentage of natural languages possessing each word order.

Embedded Relative Clause: *fruits ga which John ga said that pasta ga nibbles received wall o*

In the Recursive test set, two relative clauses are nested. In the Embedded test set, a relative clause is embedded in a subordinate clause, e.g., "he said". Through randomly sampling lexical items for the above two fixed construction templates, 500 examples for each construction are generated in each of 96 word order variations. Note that the resulting sentences are longer than 8 tokens (training data). PPLs are measured in each targeted test set, and these are aggregated as the TA score, correlation between average PPL over three model with different random seeds and word order frequency across 96 languages.

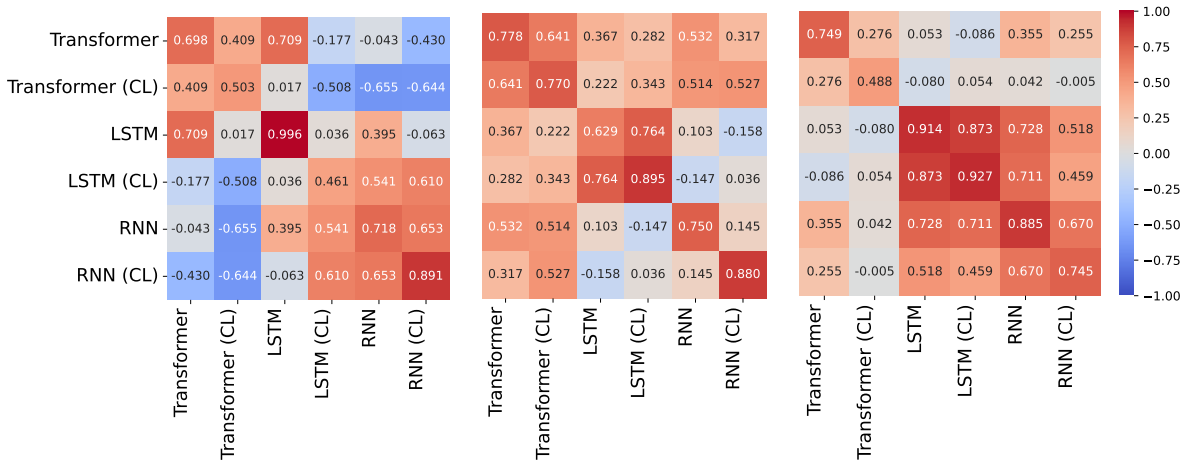
Table 2 (left-side) shows results with and without CL. CL typically yields comparable or slightly worse TA scores, while the relative order of TA scores among conditions is basically preserved, e.g., Transformer in Embedded yields the lowest TA, while RNN in Recursive yields the highest TA with or without CL.

4.2.2. Grammaticality Judgments.

We also perform grammaticality judgment evaluation with two test suites: (i) case-type accuracy, and (ii) verb-type accuracy. English-like (`0101101`) examples are as follows:

Case-type judgment: [4]

[4] *ga* and *o* in the examples are subject and object case



(a) SHORT (b) MEDIUM (c) LONG

Figure 3: Correlation of word order preference between different models×CL. Transformer (CL), for example, denotes Transformer-based LM trained with CL. Orthogonal elements denote seed variance, i.e., average correlations between the same model but different seeds.

		RECURSIVE	EMBEDDED	Case Judgment		Verb Judgment	
Model	CL	TA (↓)	TA (↓)	Acc. (↑)	Corr. (↑)	Acc. (↑)	Corr. (↑)
Trans. (El-Naggar et al., 2025b)		−5.1	−23.5†	97.7±1.5	0.14	81.0±14.7	0.27†
Trans.	✓	5.5	−19.2†	96.3±3.1	0.01	74.4±22.2	0.26†
LSTM (El-Naggar et al., 2025b)		9.2	−3.7	97.2±1.4	0.03	85.1±9.6	0.28†
LSTM	✓	16.2	0.4	95.6±1.9	-0.31†	78.1±11.7	0.32†
RNN (El-Naggar et al., 2025b)		12.9	−18.1†	97.4±1.4	0.21†	77.4±15.5	0.23†
RNN	✓	17.0	−18.9†	92.6±5.8	0.10†	63.0±25.1	0.31†

Table 2: Correlation between PPLs and word-order frequencies in targeted generalization sets (Recursive/Embedded) and correlation between accuracy and word-order frequencies in grammaticality judgment tests. Statistical significance ($p<0.05$) is marked with †.

- *fluffy soft and intelligent mango ga controls owl o*
- **fluffy soft and intelligent mango o controls owl o*

Verb-type judgment: [5]

- *scooter ga which green machine ga escorts walk*
- **scooter ga which green machine ga evolves walk*

For the case-type judgment, a randomly selected grammatical case marker is replaced with an ungrammatical one, e.g. "ga" is replaced with "o" or vice versa. For verb-type judgment, a transitive verb in a grammatical sentence is replaced with an intransitive verb, resulting in the sentence becoming ungrammatical (marked with * in the examples above). For each word order, 500 items are sampled from the MEDIUM test set, and their ungrammatical variants are created by applying specific modifications. Should there be more than one potential replacement candidate in a sentence, one is chosen at random. We report accuracy (Acc.) of grammatical judgments and correlation (Corr.) with typological commonality — the higher the better.

Table 2 (right side) shows the results. CL consistently worsened the accuracies, and the typological alignments are also weakened (Case-type judgment) or comparable (Verb-type judgment). The non-positive effect of CL is similar to the other experiments.

5. Discussion

Overall, CL led to a slightly negative effect for typological alignment between learnability and typological patterns, which leads to some interpretations. First, the results empirically demonstrated the considerable interaction effects between LMs' word-

markers, respectively.

[5] *escorts* and *evolves* in the examples are transitive and intransitive verbs, respectively.

order preferences, or more generally, inductive biases, and how data are presented. It is worth revisiting LMs' learning biases in developmentally motivated language-learning scenarios. Second, if the length-based CL we adopted is truly a good proxy for human language acquisition, the alignment between LMs' inductive biases and typological patterns might have been overestimated in existing non-CL studies. This view, more or less, challenges the idea that learners' biases shape language design and, indirectly, evokes the dominance of alternative hypotheses associating language design with cognitive-external factors, e.g., historical accidents or geographical influence (Moravcsik, 1978; Atkinson, 2011); see Culbertson et al. (2012) i.a. for a more comprehensive overview. Third, otherwise, if we stand on the premise of the learner's bias shaping language design, the weaker typological alignment under the adopted length-based CL raises the concerns that the length-based control is too simplified CL strategy, given broader possibility of CL implementations (Wang et al., 2022), or CL alone is insufficient to simulate the plausible biases in language acquisition (Perfors, 2012). These views at least poses the next question: are there any other CL strategies or factors in addition to CL leading to better simulating the emergence of typological patterns?

6. Conclusion

In this paper, we revisit the prior work (El-Naggar et al., 2025b) by introducing one basic type of curriculum learning (CL). In our experiments, we ablate the CL effect on the inductive bias of RNN, LSTM, and Transformer LMs towards different word-order configurations. We confirm that the LMs' word-order preferences indeed change under CL, suggesting that the order of the training data affects apparent LM learning biases. Our results show subtle and somewhat inconsistent interactions between individual ALs, model architectures, learning scenarios, and inductive bias. In the future, we can explore several directions to extend this line of work and gain further insights. One of them is to investigate different CL strategies, such as the construction-complexity approach or model-based control of working memory growth (Wang et al., 2022; Mita et al., 2025).

Limitations

In addition to the future directions mentioned in § 6, the definition of typological alignment should be carefully addressed. For example, we have used Pearson's correlation to examine the relationship between typological commonality and perplexity in this study, following Kuribayashi et al. (2024) and El-Naggar et al. (2025b), but its validity should be re-investigated. Finally, this study explanatorily analyzed the effect of CL on LMs' inductive bias evaluation and does not derive any strong conclusion like CL clearly promotes/hinders typological alignment between LMs' learning biases and typological commonalities. We rather hope that our paper introduces and encourages this interdisciplinary exploration space combining LMs' training scenarios and typological alignment, particularly given the progressive interdisciplinary integration between language modeling and language science.

Ethics Statement

The data used in this paper is all artificial data based mostly on English words. We have no ethical concerns with the contents of this paper.

Acknowledgment

We would like to thank the anonymous reviewers for their insightful feedback.

References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Quentin D Atkinson. 2011. Phonemic diversity supports a serial founder effect model of language expansion from africa. *Science*, 332(6027):346–349.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Nadav Borenstein, Anej Svete, Robin Chan, Josef Valvoda, Franz Nowak, Isabelle Augenstein, Eleanor Chodroff, and Ryan Cotterell. 2024. What languages are easy to language-model? a perspective from learning probabilistic regular languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15115–15134, Bangkok, Thailand. Association for Computational Linguistics.

Ted Briscoe. 1997. Co-evolution of language and of the language acquisition device. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 418–427, Madrid, Spain. Association for Computational Linguistics.

Ted Briscoe. 2000. Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, 76(2):245–296.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*, 8.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition*, 122(3):306–329.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. 2023. Neural networks and the chomsky hierarchy. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. 2025a. Gcg-based artificial languages for evaluating inductive biases of neural language models. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 540–556.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. 2025b. Which word orders facilitate length generalization in LMs? an investigation with GCG-based artificial languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35599–35613, Suzhou, China. Association for Computational Linguistics.

Nadine El-Naggar, Pranava Madhyastha, and Tillman Weyde. 2022. Exploring the long-term generalization of counting behavior in RNNs. In *I Can't Believe It's Not Better Workshop: Understanding Deep Learning Through Empirical Falsification*.

Nadine El-Naggar, Pranava Madhyastha, and Tillman Weyde. 2023. Theoretical conditions and empirical failure of bracket counting on long sequences with linear recurrent networks. *EACL 2023*, page 143.

Jeffrey L Elman. 1991. Incremental learning, or the importance of starting small. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 13.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Carla L Hudson Kam and Elissa L Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Lang. Learn. Dev.*, 1(2):151–195.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14691–14714. Association for Computational Linguistics.

Carla L Hudson Kam and Elissa L Newport. 2009. Getting it right by getting it wrong: when learners change languages. *Cogn. Psychol.*, 59(1):30–66.

Kai A Krueger and Peter Dayan. 2009. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent word order universals from cognitively-motivated language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14522–14543. Association for Computational Linguistics.

S. J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume*

1: Long Papers, pages 4975–4989. Association for Computational Linguistics.

Masato Mita, Ryo Yoshida, and Yohei Oseki. 2025. Developmentally-plausible working memory shapes a critical period for language acquisition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9386–9399, Vienna, Austria. Association for Computational Linguistics.

Edith Moravcsik. 1978. Language contact. *Universals of human language*, 1:93–122.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Amy Perfors. 2012. When do memory limitations lead to regularization? an experimental and computational investigation. *J. Mem. Lang.*, 67(4):486–506.

Douglas LT Rohde and David C Plaut. 1999. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109.

Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2024. Targeted syntactic evaluation on the chomsky hierarchy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 15595–15605. ELRA and ICCL.

Valentin I Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2009. Baby steps: How "less is more" in unsupervised dependency parsing. *NIPS: Grammar Induction, Representation of Language and Language Learning*, pages 1–10.

Mark Steedman. 1996. *Surface structure and interpretation*, volume 30 of *Linguistic inquiry*. MIT Press.

Mirac Suzgun, Sebastian Gehrmann, Yonatan Belinkov, and Stuart M Shieber. 2019. Lstm networks can perform dynamic counting. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pages 44–54.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A survey on curriculum learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4555–4576.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press, Boca Raton.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision RNNs for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 740–745. Association for Computational Linguistics.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 454–463. Association for Computational Linguistics.

Mary McGee Wood. 2014. *Categorial grammars (RLE linguistics b: Grammar)*. Routledge.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. 2025. Can language models learn typologically implausible languages?

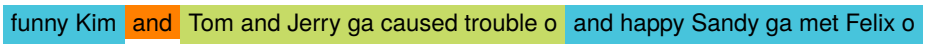
A. Categorial Grammar

Categorial Grammar (CG) is a formalism consisting of a lexicon, where each word is assigned a functor category, and a set of rules that define how the categories combine with each other. CG was introduced with the combinatory operation **application**, which has forward and backward variants. Extensions of CG have been introduced, such as combinatory categorial grammar (CCG) (Steedman, 1996) and generalized categorial grammar (GCG) (Wood, 2014). El-Naggar et al. (2025b) use a GCG with the following combinatory operations, which they detail in their paper:

- Application (forward and backward variants)
- Coordination
- Composition (forward and backward variants)
- Weak Generalized Permutation

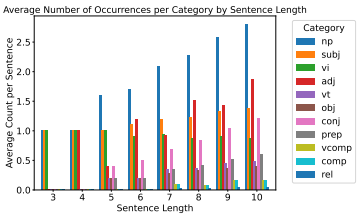
This GCG allows for the expression of mildly context-sensitive constructions, like cross-serial and unbounded dependencies (Briscoe, 1997, 2000). Furthermore, by using GCG-based ALs in experiments, we are able to test a wider attested range of word order configurations including VSO and OSV base word order ALs.

(a) Example of two valid sentences are concatenated with a conjunction to create a longer one.

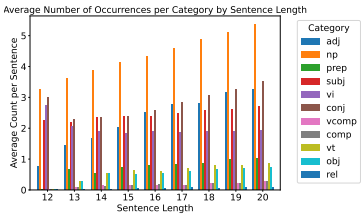


(b) Example of how a valid sentence is embedded into another valid sentence to create a longer one.

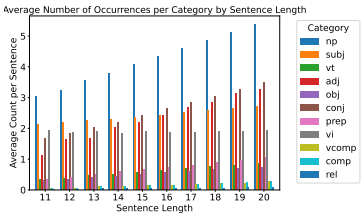
Figure 4: Examples of short templates (lengths 3-10) being combined to create longer templates (lengths 11-20). The first sentence is in blue, the second sentence is in green, and the conjunction is in orange.



(a) How many times each category appears in SHORT and MEDIUM per template on average.

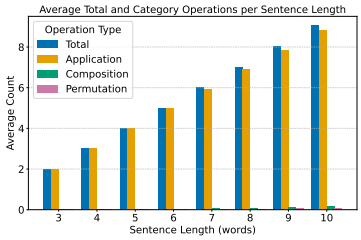


(b) How many times each category appears in templates used to create LONG set, per template on average.

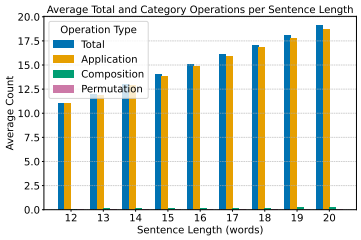


(c) How many times each category appears in LONG per template on average.

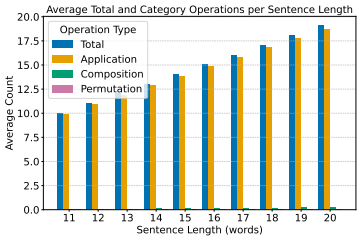
Figure 5: The number of occurrences of the different categories for all templates lengths 3-10 (a), the templates used to create the long dataset (b), and all the generated long templates (c).



(a) Template lengths 3-10 words (SHORT and MEDIUM).



(b) Sampled templates used to create the LONG dataset.



(c) All extended templates created of lengths 11-20.

Figure 6: The average number of combinatory operations in the GCG derivation for the templates of different lengths for the language configuration `0101101` by El-Naggar et al. (2025b).

B. Dataset Details

B.1. Dataset Generation

The ORIGINAL, SHORT, MEDIUM and LONG datasets are the same ones used by El-Naggar et al. (2025b).

- **SHORT Dataset:** The sentences in this dataset are created by generating all possible category combinations, which are referred to as templates, of length 3 to 8 words. The templates are then parsed and assigned to the ALs where they would produce a valid parse, i.e. result in an S node spanning the input. All lengths are represented equally in the dataset, and within each length, the different templates are also represented equally.

- **MEDIUM Dataset:** Like the SHORT dataset, this dataset is created by generating all possible templates of lengths 9-10 words. The templates are then parsed and assigned to ALs where they produce a valid parse. Similarly, all lengths are represented equally and within each length all templates are represented equally. The lexicon is then randomly sampled to create unique sentences.

- **LONG Dataset:** To generate the LONG dataset, the templates for the SHORT and MEDIUM datasets are combined in 3 ways, as shown in Figure 4:

  1. Appended end to end
  2. Concatenated with a conjunction (Fig. 4a),
  3. Embedded with a conjunction (Fig. 4b).

  The resulting longer templates are parsed to filter out ungrammatical ones. Because there are millions of valid templates of length 11-20, 20K templates are randomly sampled, and for each one, the lexicon is sampled. It is worth

Fairseq model	share-decoder-input-output-embed	True
	embed_dim	128
	ffn_embed_dim	512
	layers	2
	heads	2
	dropout	0.3
	attention_dropout	0.1
	#params.	462K
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	400
	warmup init learning rate	1e-7
Training	batch size	2,048 tokens
	tokens-per-sample	128 tokens
	sample-break-mode	none
	epochs	10

(a) Transformer.

Fairseq model	share-decoder-input-output-embed	True
	embed_dim	128
	hiden_size	512
	layers	2
	dropout	0.1
	#params.	3,547K
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	400
	warmup init learning rate	1e-7
Training	batch size	2,048 tokens
	tokens-per-sample	128 tokens
	sample-break-mode	none
	epochs	10

(b) LSTM.

Fairseq model	share-decoder-input-output-embed	True
	embed_dim	64
	hiden_size	64
	layers	2
	dropout	0.1
	#params.	49K
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	400
	warmup init learning rate	1e-7
Training	batch size	2,048 tokens
	tokens-per-sample	128 tokens
	sample-break-mode	none
	epochs	10

(c) RNN.

Table 3: Hyperparameters of LMs.

noting that for the English-like language configuration `0101101`, the vast majority of the valid extended templates will have been created by concatenation with a conjunction (b) or embedding with a conjunction (c). However, it may be possible for the end to end appending of templates to result in valid templates in other language configurations.

B.2. Dataset Statistics

We show the occurrences of the different categories in the templates in El-Naggar et al. (2025b) for lengths 3-10 in Figure 5a. The occurrences of the templates used to create the LONG dataset are shown in Figure 5b, and in Figure 5c for all generated templates of length 11-20. We can see in Figure 5 that the number of categories that appear in shorter sentences is smaller than in longer sentences, and that the number of category occurrences increases with sentence length. We show in Figure 6 the average number of different combinatory operations that are required to derive the templates of different lengths.

C. Model Hyperparameters

We use exactly the same model hyperparameters as El-Naggar et al. (2025b). These are summarised in Table 3.