Evaluating Neural Language Models as a Cognitive Model of Human Second Language Acquisition by Comparing the Developmental Trajectory Between Human and Model Learning

Rin Otokawa, Tomoki Miyamoto, and Akira Utsumi

1. Introduction

Recent advances in neural language models have sparked growing interest in how language models (LMs) acquire language. In the domain of first language (L1) acquisition modeling, Evanson et al. (2023) trained LMs (GPT-2) from an initial state with no prior knowledge. They tracked the process and compared it with child language acquisition data. Their results showed that the model acquired language skills in a consistent order, and that this order matched the developmental stages of children. Many other studies have also been conducted on the cognitive models of L1 acquisition (Warstadt et al., 2023; Goldstein et al., 2022).

In addition to studies on L1 acquisition, researchers have also begun to explore how LMs acquire second languages (L2). When testing whether LMs are plausible as a cognitive model of L2 acquisition, a key issue is whether the major phenomena observed in human language acquisition can be simulated in the LMs. In human L2 acquisition, a phenomenon known as "cross-linguistic transfer" exists, where L2 learning is facilitated or inhibited by the influence of the L1. Many empirical studies have demonstrated this phenomenon (Schepens et al., 2020; Tolentino et al. 2014).

Yadavalli et al. (2023) investigated cross-linguistic transfer in L2 (English) acquisition using a model (SLABERT) pre-trained on L1 child-directed speech data. Using the Grammaticality Judgment Benchmark (BLiMP), they compared the final scores of models with different L1s. Their results demonstrated the presence of negative cross-linguistic transfer, in which model performance decreased as the typological distance between L1 and English increased. Oba et al. (2024) also utilized BLiMP to analyze the influence of L1 pre-training on L2 linguistic generalization. They examined the performance difference as compared to a model without pre-training and tracked developmental curves. Their results revealed that L1 knowledge generally facilitated L2 learning, but the extent of this effect varies significantly depending on linguistic phenomenon in question.

*Rin Otokawa, The University of Electro-Communications, o2430026@edu.cc.uec.ac.jp; Tomoki Miyamoto, The University of Electro-Communications, miyamoto@uec.ac.jp; Akira Utsumi, The University of Electro-Communications, utsumi@uec.ac.jp.

 Proceedings of the 50th annual Boston University Conference on Language Development, ed. Romi Hill et al., 600-612. Somerville, MA: Cascadilla Press.

However, although these studies demonstrated the learning capabilities and cross-linguistic transfer of the LMs, challenges remain in testing the validity as cognitive models. Specifically, whether the L2 acquisition process of L1-trained LMs is similar to a human-like developmental trajectory has not been sufficiently examined. Yadavalli et al. focused exclusively on models at a single point in time after the completion of L1 and L2 training. Although Oba et al. analyzed developmental curves using BLiMP, this benchmark is designed to evaluate general English grammatical competence and does not reflect the specific error patterns or developmental stages unique to L2 learners. Thus, their evaluation did not test whether the model trajectory mimics the actual human-like process of L2 acquisition.

In this study, we test the validity of LMs as cognitive models of L2 acquisition by focusing on negative cross-linguistic transfer in grammatical knowledge and comparing the L2 acquisition processes of humans and LMs. Specifically, we examine whether the LMs internalize the error patterns exhibited by human L2 learners at different proficiency levels and whether it can simulate the trajectory of overcoming these errors as learning progresses. The contribution of this study lies in the evaluation of the "human-likeness of the developmental trajectory of LMs," which has received little attention in previous research.

2. Cognitive Model

2.1. Language Model

In this study, we used the XLM (Lample et al., 2019), which was used by Oba et al. (2024), as the LM for evaluation. XLM is a model based on the Transformer encoder (Vaswani et al., 2017). The model was trained using two methods: Masked Language Modeling (MLM) and Translation Language Modeling (TLM). MLM is designed to learn monolingual representations. TLM is designed to acquire L2 representations by using the L1 knowledge learned through MLM. In MLM, tokens in the input text are randomly masked, and the model learns to predict them from the bidirectional context of the entire sentence. In TLM, tokens in concatenated parallel sentence pairs are randomly masked, and the model learns to predict them. In TLM, when predicting a masked token, the model can refer to contextual information not only within the same language, but also from the other concatenated language. This enables the learning of multiple languages within a shared semantic space.

2.2. Construction of the Cognitive Model

In this study, we constructed two XLM-based models. The first model is a JA-EN model that intended to simulate a native Japanese speaker learning English as an L2. The second model is an EN-Only model intended to simulate a native

English speaker for comparison. In this section, we describe the dataset used for training, preprocessing, and the methods for constructing cognitive models.

2.2.1. Dataset and Preprocessing

For model training, we used a Japanese monolingual corpus and a Japanese-English parallel corpus. As data for L1 (Japanese) pre-training, we extracted and used the initial 5.3 million lines from the Japanese data of the CC100 corpus (Conneau et al., 2020). This text was first tokenized using the morphological analyzer MeCab (Kudo et al. 2002), and subsequently segmented into subword units using fastBPE (Sennrich et al. 2016). We trained a subword segmenter with a vocabulary size of 14,000, and the dataset was split into training, validation, and test sets at a ratio of 98:1:1.

For the Japanese-English parallel data, we used approximately 280,000 parallel sentence pairs extracted from the Tatoeba corpus (Tiedemann et al., 2012). The Japanese sentences were tokenized with MeCab, and the English sentences were processed with the Moses toolkit tokenizer (Koehn et al., 2007). FastBPE was then applied to each tokenized dataset for subword segmentation. The same merge rules from the L1 training were applied to the Japanese data, whereas a separate subword segmenter with a vocabulary size of 6,000 was trained for the English data. From the parallel pairs, the dataset was split into training, validation, and test sets at a ratio of approximately 99:0.5:0.5.

Furthermore, to create an English monolingual corpus, we extracted the English side of the Tatoeba corpus.

2.2.2. Model Construction Procedure

The JA-EN and EN-Only models were trained using the datasets described in Section 2.2.1 according to the following procedure.

The JA-EN model was constructed in two stages: L1 acquisition and L2 acquisition. In the first stage (L1 acquisition), an initialized XLM was pre-trained on the Japanese monolingual corpus using a Masked Language Modeling (MLM) objective for 10 epochs. In the second stage (L2 acquisition), the L1-trained model underwent an additional 180 epochs of training using the Japanese-English parallel corpus and the English monolingual corpus. To maintain both translation ability and English fluency, we alternated training tasks with each epoch. Specifically, Translation Language Modeling (TLM) and MLM (using English monolingual data) were applied alternately. Before starting the second stage, the weights and biases of the embeddings and output layers were adjusted to correspond to the expanded vocabulary size.

For comparison, the EN-Only model was constructed using only the English monolingual corpus. This model was trained using an initialized XLM with MLM.

To ensure the reliability of the evaluation results, we conducted three independent training runs with different random seeds. Unless otherwise specified,

the mean accuracy across these three models is reported in the subsequent analysis.

2.2.3. Training Parameter Settings

The XLM models, comprising approximately 18M parameters, were trained using the hyperparameters in Table 1.

Table 1. Hyperparameters for XLM training

Category	Parameter	Value
Model Architecture	Dropout rate	0.1
	Attention dropout	0.1
	Embedding/Hidden dimension	256
	Feed-forward dimension	1,024
	Number of layers	12
	Number of heads	8
	Activation function	GELU
Optimization	Algorithm	Adam
	Learning rate	0.0002
	β_1, β_2	0.9, 0.999
	ε	10^{-6}
	Weight decay	0.01
	Scheduler	Inverse sqrt
	Warmup steps	30,000
Training Settings	Gradient clipping (norm)	1.0
	Gradient accumulation	4
	FP16 training	Yes (amp=2)

3. Method

In human L2 acquisition of grammatical knowledge, the influence of negative cross-linguistic transfer varies according to the learner's proficiency level. For example, Japanese learners of English at the beginner level tend to omit articles because the grammatical category of articles does not exist in Japanese, which leads to errors such as "I am student." Although these types of elementary errors tend to decrease as proficiency increases, other types of errors may emerge as learners attempt to use newly acquired expressions.

In this study, we evaluated the influence of negative cross-linguistic transfer on the acquisition of grammatical knowledge in LMs using a dataset that reflects the specific error patterns of Japanese learners of English at different proficiency levels.

3.1. Dataset Construction

To construct a dataset reflecting the error patterns of L2 learners by proficiency level, we used the NICT JLE Corpus (Tono et al. 2005), which contains spoken English produced by Japanese learners. In this Corpus, learners were categorized into nine proficiency levels ranging from Level 1 to 9 according to their English proficiency. In this study, however, we grouped these levels into three broader categories: "Beginner (Group 1-3)", "Intermediate (Group 4-6)", and "Advanced (Group 7-9)".

In the NICT JLE Corpus, a total of 47 types of error tags combining "Part of Speech (Level 1)" and "Nature of Error (Level 2)" are assigned to learner errors. Table 2 shows some error tags specifically used in this study and their examples of error sentences.

Table 2. Error tags and their examples in the NICT JLE Corpus

Tag	Description	Example (Incorrect → Correct)	Group
`n_num`	Noun: Number	I have three **brother**. → I have three **brothers**.	Group 2
`v_tns`	Verb: Tense	And it **isn't** really anything special, but I just enjoyed that atmosphere. → And it **wasn't** really anything special, but I just enjoyed that atmosphere.	Group 9
`v_agr`	Verb: Agreement	My wife **go** to movies with him. → My wife **goes** to movies with him.	Group 4

In this study, we created error-correction pairs by extracting sentences with tags and labeling the original erroneous sentences as 'incorrect' and their corrected versions as 'correct'. The resulting dataset comprises 1,603 pairs for the beginner level (Groups 1–3), 6,872 pairs for intermediate level (Groups 4–6), and 1,853 pairs for advanced level (Groups 7–9). During the collection process, we excluded utterances that were structurally fragmented or unintelligible (e.g., those tagged as o_uit). Furthermore, to ensure the quality of the sentence pairs, we filtered the pairs according to the following three steps.

1. **Noise Removal**: Superficial errors, such as capitalization and minor spelling inconsistencies unrelated to the acquisition of linguistic knowledge targeted in this study, were excluded.

2. **Selection by Multiple Pretrained Language Models**: Using three language models with different characteristics (RoBERTa-base, GPT2-medium, XLNet-base-cased), we adopted only the sentence pairs that two or more of these models judged correctly.

3. **Removal of Sentences Related to Semantic Knowledge**: For pairs with lexical error tags (`n_lxc`, `v_lxc`, etc.), we compared the error location and the lemma of the corrected word using spaCy (Honnibal et al., 2020). We categorized pairs with the same lemma (e.g., *go* and *went*) as "grammatical knowledge" and pairs with different lemmas (e.g., *high* and *tall*) as "semantic knowledge," and excluded pairs categorized into semantic knowledge.

The number of filtered sentence pairs at each filtering step and the final size of the evaluation dataset are in Table 3.

Table 3. Number of filtered sentence pairs at each filtering step and the final size of the evaluation dataset

Proficiency Group	Initial Count	Step 1	Step 2	Step 3	Final Count
Beginner (Groups 1–3)	1,603	182	250	129	1,042
Intermediate (Groups 4–6)	6,872	660	1,096	620	4,496
Advanced (Groups 7–9)	1,853	157	358	169	1,169

3.2. Evaluation Method

LMs are trained in epochs, which represent one complete pass through all of the training data. At each epoch during English learning, the LMs judged which of the sentences in a pair in Section 3.1 is correct. For this judgment, we used Pseudo-Perplexity (PPPL) (Salazar et al., 2020). PPPL is an indicator of how accurately tokens at each position can be inferred from the surrounding context and is defined by the following equation:

$$\mathrm{PPPL}(s) = \left(\prod_{t=1}^{n} p_\theta(w_t | s_{\setminus w_t}) \right)^{-\frac{1}{n}} \tag{1}$$

where $s_{\setminus w_t}$ represents the remaining context after removing (masking) the t-th token w_t from sentence s, and $p_\theta(w_t|s_{\setminus w_t})$ is the probability of predicting the original token w_t from that context. Correct sentences tend to have lower PPPL because the LMs predict constituent words with higher probability. Hence, the sentence with the lower PPPL is reasonably assumed to be judged as correct by the model, and the proportion of cases where the model's judgment matched the actual correction was calculated as "accuracy."

To evaluate whether the model mimics the human L2 developmental trajectory, we observed the transition of accuracy by proficiency level. Figure 1 shows the ideal developmental curve of human L2 acquisition. For English learners whose native language is Japanese, the difficulty of the sentence pairs in the

dataset presented in Section 3.1 is considered to increase in the order of beginner, intermediate, and advanced. Therefore, it is assumed that knowledge at the beginner level is acquired first, followed by knowledge at the intermediate level, and finally, knowledge at the advanced level. Ultimately, the accuracy across all proficiency groups is expected to converge to the same level, reflecting the final stage of human language acquisition. Thus, if the JA-EN model shows a similar developmental curve, it is a plausible cognitive model of L2 acquisition. Furthermore, to rule out the possibility that an LM without L1 learning could also exhibit a developmental curve similar to the JA-EN model, we conducted the same evaluation on the EN-Only model. If the EN-Only model does not exhibit the same developmental curve, then the behavior of the JA-EN model is confirmed to be attributed to L1 knowledge.

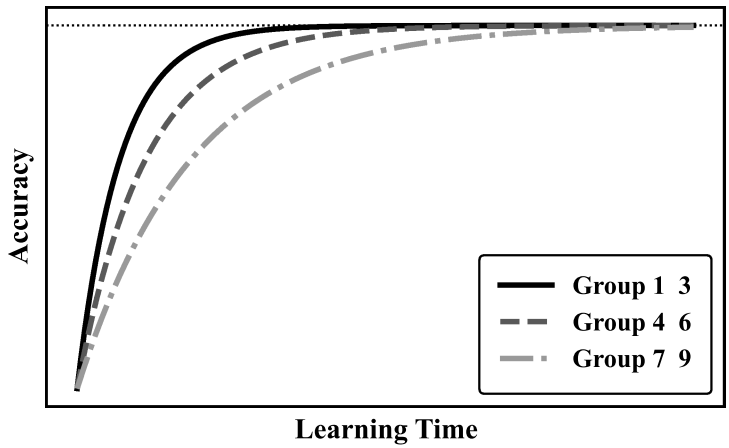
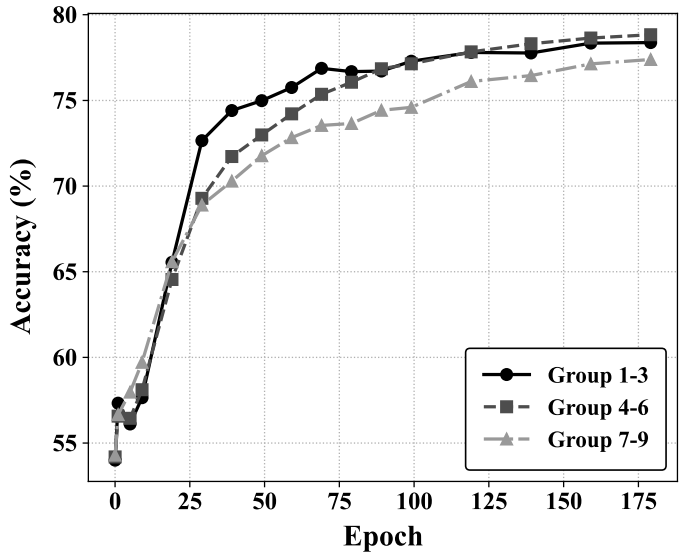


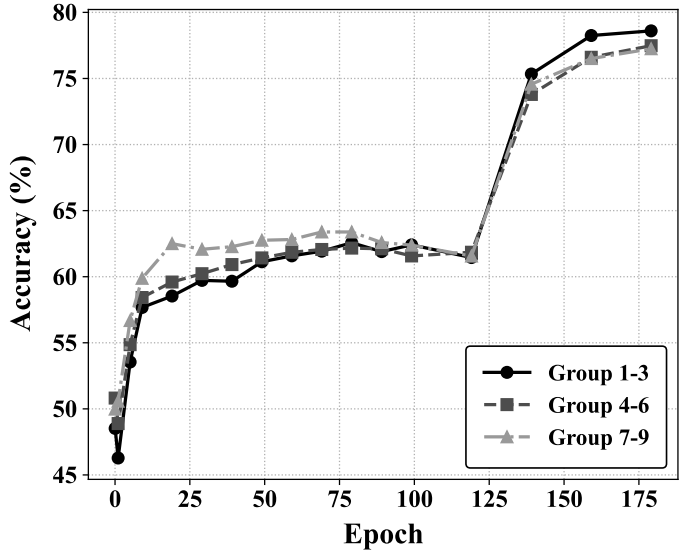
Figure 1. Ideal developmental trajectory in L2 acquisition representing different proficiency groups

4. Results

The accuracy of the two XLM models for each epoch on the sentence pairs created in Section 3.1 are shown in Figure 2(a) and 2(b).



(a) JA-EN Model



(b) EN-Only Model

Figure 2. Developmental curve of grammar accuracy for (a) JA-EN and (b) EN-Only models

In the JA-EN model, the accuracy for the beginner level (Group 1-3) improved rapidly from the beginning of training, reaching about 77% around epoch 70. No further improvement was observed after this point. Next, the intermediate-level (Group 4-6) accuracy improved more gradually than Group 1-3, converging to about 77% around epoch 100. Finally, the advanced-level (Group 7-9) accuracy improved most gradually, converging to about 77% accuracy, almost equivalent to the other groups, around epoch 150. Thus, a time lag is observed in the improvement of accuracy for each group, indicating that errors are overcome in stages according to proficiency. This is consistent with the ideal developmental trajectory of human L2 acquisition shown in Figure 1.

In contrast, in the EN-Only model, although the accuracy of the beginner level was low immediately after the start of training, the accuracy of each group became comparable as training progressed, and no differences in proficiency level were observed in the latter half of training. The final accuracy converged to about 77% for all groups.

The JA-EN model showed a human-like developmental trajectory, but the EN-Only model did not. This demonstrates the validity of the JA-EN model as a cognitive model of L2 acquisition.

5. Discussion

5.1. Validity as a Cognitive Model

The sentence pairs used in this study reflect the error patterns of human L2 learners across proficiency level. For the JA-EN model, its developmental curve of overcoming errors was similar to the ideal developmental trajectory in human L2 acquisition. On the other hand, no such tendency was observed in the EN-Only model. This indicates that the JA-EN model was influenced by its L1 knowledge and internalized the error patterns observed in human L2 acquisition.

We conducted a Friedman test to verify the statistical significance of accuracy differences in the JA-EN model. This analysis compared proficiency groups across the 180-epoch L2 acquisition phase, followed by the Nemenyi method for post-hoc pairwise comparisons. The results of these tests are shown in Table 4.

Table 4. Results of Friedman tests and Nemenyi post-hoc tests

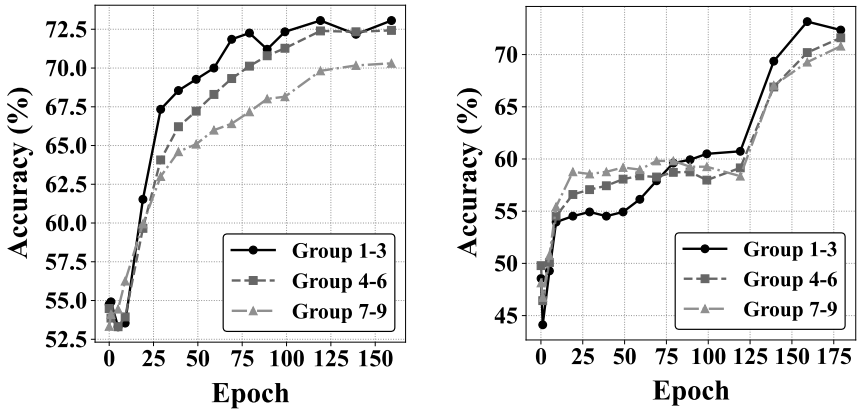
		Median Accuracy			
Condition / Trial	p-value	G1-3	G4-6	G7-9	Significant Pairs (Nemenyi)
JA-EN Model					
Average	0.056	**0.758**	0.742	0.728	*n.s.*
Trial 1	0.002^{**}	**0.757**	0.739	0.727	G1-3 > G7-9**, G4-6 > G7-9**
Trial 2	0.101	**0.754**	0.745	0.726	*n.s.*
Trial 3	0.005^{**}	**0.761**	0.742	0.731	G1-3 > G7-9**

Note: G1-3: Beginner, G4-6: Intermediate, G7-9: Advanced.
Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

First, the test on the mean accuracy of the JA-EN model resulted in $p = 0.056$, which is marginally above the 0.05 significance level. The results indicated a marginal trend toward differences in accuracy among the three proficiency levels. To further examine whether such a trend can be observed, we analyzed the results of each trial separately. Significant differences were confirmed in Trials 1 and 3 ($p < 0.005$). More importantly, the median accuracy consistently followed the hierarchy of G1-3 > G4-6 > G7-9 in every single trial, including Trial 2. The post-hoc analyses for the significant trials further confirmed that accuracy at the beginner (G1-3) and intermediate (G4-6) levels significantly exceeded that of the advanced level (G7-9). The fact that this specific proficiency hierarchy (Beginner > Intermediate > Advanced) was consistently observed across all independent runs indicates that the result is stable and not due to random chance. This supports the conclusion that the JA-EN model is influenced by negative transfer from Japanese (L1) and internalizes the same error patterns and acquisition order as human learners.

5.2. Effectiveness and Limitations of Filtering the NICT JLE Corpus using Language Models

To verify the effectiveness of the filtering process, we re-evaluated the XLM models using the dataset without filtering, as shown in Figure 3. When the unprocessed dataset was used, the final overall accuracy was consistently lower compared to the filtered dataset (Figure 2). Specifically, for the EN-Only model, the final accuracy of all groups decreased from approximately 77% (Figure 2(b)) to around 70% (Figure 3(b)). For the JA-EN model, the final overall accuracy decreased from approximately 77% (Figure 2(a)) to about 72.5% for the beginner and intermediate levels, and to around 70% for the advanced level (Figure 3(a)). These results suggest that the filtering procedure contributed to reducing noise in the dataset.



(a) JA-EN Model

(b) EN-Only Model

Figure 3. Developmental curves of grammar accuracy for (a) JA-EN and (b) EN-Only models without model-based filtering

However, the validity of the filtering procedure is debatable. Since language models were used for data selection, the resulting filtered dataset may be biased toward linguistic patterns preferred by the models. This potential bias should be considered a limitation of this study.

6. Conclusion

In this study, we constructed a cognitive model of English L2 learners with Japanese as L1 using XLM and tested its validity. The JA-EN model showed a developmental curve consistent with the ideal developmental trajectory in human L2 acquisition. However, the EN-Only model, which lacked L1 knowledge, did not show this same developmental curve. Therefore, our findings suggest that the XLM's JA-EN model is plausible as a cognitive model of human L2 acquisition.

In future work, whereas this study focused on the acquisition of grammatical knowledge, a comprehensive evaluation regarding the acquisition of semantic knowledge is also necessary. Furthermore, although only XLM was used in this study, it is necessary to examine other model architectures to more generally evaluate the validity of LMs as cognitive models of L2 acquisition.

References

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Evanson, Linnea, Yair Lakretz & Jean-Rémi King (2023). Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics.

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, David Emanuel, Alon Cohen & others (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25:369–380.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Adriane Boyd (2020). spacy: Industrial-strength natural language processing in Python. Zenodo.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin & Evan Herbst (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kudo, Taku & Yuji Matsumoto (2002). Japanese dependency analysis using cascaded chunking. *IPSJ Journal*, 43(6):1834–1842.

Lample, Guillaume & Alexis Conneau (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32, pages 7057–7067.

Oba, Miyu, Tatsuki Kuribayashi, Hiroki Ouchi & Taro Watanabe (2024). Second language acquisition in language models. *Journal of Natural Language Processing*, 31(2):433–455.

Salazar, Julian, Davis Liang, Toan Q. Nguyen & Katrin Kirchhoff (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Schepens, Job, Roeland van Hout & T. Florian Jaeger (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, 194:104056.

Sennrich, Rico, Barry Haddow & Alexandra Birch (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Tiedemann, Jörg (2012). Parallel data, tools and interfaces in OPUS. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis, editor, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Tolentino, Leida C. & Natasha Tokowicz (2014). Cross-language similarity modulates effectiveness of second language grammar instruction. *Language Learning*, 64(2):279–309.

Tono, Yukio, Emi Izumi & Emiko Kaneko (2005). The NICT JLE corpus: The final report. In *Proceedings of the JALT2004 Conference*, pages 345–356.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Warstadt, Alex, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Cheng, Chengxu Zhuang, Jennifer Hu & others (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 1–34.

Yadavalli, Aditya, Alekhya Yadavalli & Vera Tobin (2023). SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777.

Proceedings of the 50th annual Boston University Conference on Language Development

edited by Romi Hill, Xinhan Jiang, Danutham Worapipat, and Aditya Yedetore

Cascadilla Press Somerville, MA 2026

ISSN 1080-692X
ISBN 978-1-57473-047-0 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com