Second Language Acquisition of Neural Language Models

Miyu Oba[1] **Tatsuki Kuribayashi**[2,3] **Hiroki Ouchi**[1,4] **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology
[2]MBZUAI [3]Tohoku University [4]RIKEN
{oba.miyu.ol2, hiroki.ouchi, taro}@is.naist.jp
tatsuki.kuribayashi@mbzuai.ac.ae

Abstract

With the success of neural language models (LMs), their language acquisition has gained much attention. This work sheds light on the **second language (L2) acquisition** of LMs, while previous work has typically explored their first language (L1) acquisition. Specifically, we trained bilingual LMs with a scenario similar to human L2 acquisition and analyzed their cross-lingual transfer from linguistic perspectives. Our exploratory experiments demonstrated that the L1 pretraining accelerated their linguistic generalization in L2, and language transfer configurations (e.g., the L1 choice, and presence of parallel texts) substantially affected their generalizations. These clarify their (non-)human-like L2 acquisition in particular aspects.[1]

1 Introduction

Cross-lingual transferability of language models (LMs) has attracted much attention. For example, large English LMs show some translation performance even when using a small amount of non-English languages as training data (Brown et al., 2020; Shi et al., 2023), which indicates the efficient language transfer from English to others. Such cross-lingual transferability has been evaluated by holistic measures, such as perplexity and accuracy on downstream tasks (Papadimitriou and Jurafsky, 2020; Deshpande et al., 2022; Blevins et al., 2022). On the other hand, there is much room for investigating them from *linguistic perspectives*; e.g., grammatical knowledge acquisition and language transfer tendencies among languages.

In this study, we investigate the cross-lingual transferability of LMs from a perspective of **second language (L2) acquisition**. Our main research question is *how first language (L1) acquisition of LMs affects the efficiency of grammar acquisition in L2*. To answer this question, we design an experimental procedure (Section 2): (i) pretraining LMs in a certain language (assumed to be the L1 speakers), (ii) further training them in English as an L2, and (iii) evaluating and analyzing their linguistic generalization in L2. As L1s, we chose four languages with different levels of difficulty in transferring to English, i.e., French, German, Russian, and Japanese. The size of training data is restricted to match the human-like L2 acquisition scenario, which enables better comparison with human L2 acquisition tendencies and, hopefully, provides insights into L2 acquisition from a computational linguistic perspective.

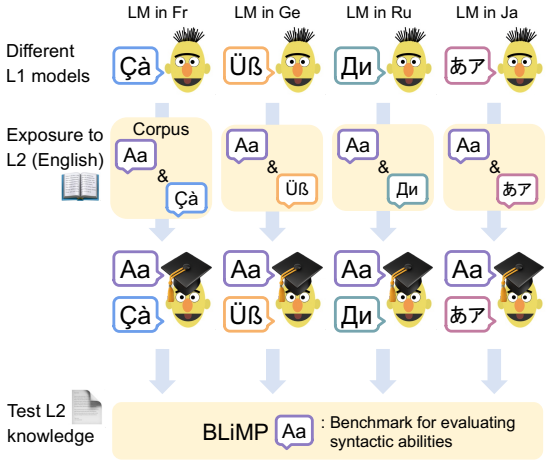

Figure 1: Experimental Procedure. First, we pretrain the monolingual masked language model on the first language (first language acquisition; L1 acquisition). Then, the model is additionally trained under the bilingual setting (second language acquisition; L2 acquisition). Finally, we analyze the effect of L1 on L2 via a grammatical judgment test in L2.

We begin with exploring the inductive biases of several L2 training methods (Section 3). Specifically, we compared some variations of L2 data settings, such as training on only the L2 texts or on L1–L2 translation pairs. We observed that, for example, feeding L1–L2 translation pairs into LMs

[1]Our codes are available at https://github.com/mlieynua/sla-of-nlm

Findings of the Association for Computational Linguistics: ACL 2023, pages 13557–13572

slowed down their L2 grammar acquisition, compared to only feeding L2 monolingual texts every two epochs.

In our main experiments, we conducted exploratory analyses of the effects of L1 training on L2 grammar acquisition (Section 4). We gained three main insights. First, L1 knowledge promotes better linguistic generalization in L2 (Section 4.1). Second, different L1s incur different generalizations in L2 (Section 4.2). More specifically, Japanese and Russian are far behind French and German, which is consistent with the human-defined difficulty levels of language transfer (Chiswick and Miller, 2004). Third, L1 pretraining gives different effects on different types of grammar items (Section 4.3). In particular, morphological and syntactic items get larger gains than semantic and syntax&semantic items.

In more detail, we analyzed the *process* of L2 acquisition (Section 5). We investigated how L2 knowledge acquisition progresses (Section 5.1) and found that L2 knowledge acquisition does not progress so much until seeing the whole dataset many times (e.g., 50-100 times), implying their data inefficiency. Furthermore, we also observed the L1 knowledge degrade during L2 acquisition; this motivates us to balance the source–target linguistic knowledge during language transfer.

2 Second language acquisition of LMs

Overview: We are interested in how L1 knowledge affects the linguistic generalization of LMs in L2. Figure 1 shows an overview of the experimental procedure. First, in our L1 acquisition simulation, we train LMs on a monolingual corpus of a specific language. Second, in our L2 acquisition simulation, we additionally train the pretrained LMs with a corpus including L2 texts (English). Finally, we evaluate the grammatical judgment ability of the LMs in the L2 (English) using BLiMP (Warstadt et al., 2020).

2.1 Language exposure

First and second languages: We used French, German, Russian, and Japanese as L1 and employed English as L2 (Table 1). We expect that the transfer to English becomes more difficult in order of French, German, Russian, and Japanese from multiple perspectives: linguistic distance (Grimes and Grimes, 2002; Chiswick and Miller, 2004) and learning difficulty level[2].

Lang.	Family	Order	Script	Rank
French	IE	SVO	Alphabet	1
German	IE	SOV	Alphabet	2
Russian	IE	SVO	Cyrillic	3
Japanese	N-IE	SOV	Kana/Kanji	4
English	IE	SVO	Alphabet	-

Table 1: Characteristics of the four languages used in our experiment. English is employed as L2, and the others are L1s. "Rank" indicates the transfer difficulty from the corresponding language to English, based on the linguistic distance and FSI rank; a higher value indicates a greater gap to English from the language acquisition perspective. "Family" indicates if Indo-European (IE) or not (N-IE). "Order" indicates canonical word order in the corresponding language.

L1 acquisition: We first train LMs in particular L1 language using a monolingual corpus of approximately 100M words sampled from CC-100 (Conneau et al., 2020; Wenzek, 2020). The corpus size is roughly similar to the number of words exposed to humans during language acquisition. We trained the models with 100 epochs.[3]

L2 acquisition: We then further train the L1 LMs under bilingual input (Section 3). We trained the models with 100 epochs, but the intermediate checkpoints will also be analyzed to track the process of L2 acquisition in Section 5. We used Tatoeba (Tiedemann, 2012)[4] as parallel corpora in L2 acquisition. Tatoeba is a multilingual parallel corpus consisting of example sentences originally collected for human foreign language learners. From the L2 acquisition perspective, this amount would be large enough for human learners to learn the top 95% English words in frequency (Nation, 2014).

Note that there would be several scenarios of human L2 learning/acquisition, such as through

[2] https://www.state.gov/foreign-language-training/ Note that these difficulty levels only indicate the difficulty of transferring from English to a specific language. In our study, we tentatively assume the symmetry of the source and target language in terms of learning difficulty.

[3] This number of epochs might be cognitively-implausible since humans would not face the same example 100 times, but it is also argued that the memory of language experience continues to affect the learning multiple times (Bybee, 2013).

[4] https://opus.nlpl.eu/Tatoeba.php

language classes or natural communications. Following Krashen et al. (1979), we refer to *L2 acquisition* as the latter scenario of acquiring L2 through natural language exposure, e.g., raw texts.

2.2 Learners

We largely followed the settings of the cross-lingual language model (XLM) (Conneau and Lample, 2019), which is typically used in cross-lingual language modeling in the field of natural language processing (NLP). In short, this is a Transformer-based bidirectional LM, but the input consists of bilingual text pairs. The tokens in the bilingual text were randomly masked, and the model predicts the masked tokens on both L1 and L2 sides. During the L1 training, the L1 side is the only input.

The bilingual XLM is trained from scratch (L1 training and L1–L2 training), rather than using the off-the-shelf pre-trained XLM that is trained across dozens of languages (Conneau and Lample, 2019; Conneau et al., 2020). From a cognitive perspective, such a super-multilingual setting is unrealistic since humans hardly face dozens of languages in a multilingual acquisition scenario. Rather, we hope that such a bilingual transfer will gain much attention from the adjacent areas, such as pedagogy and cognitive science of exploring human second language acquisition/learning.

Technically, we randomly initialized the parameters of the XLM (18M), constructed a bilingual tokenizer using byte pair encoding on the bilingual texts, and trained the model separately for each L1–L2 combination. For each L1–L2 setting, we trained four models with different seeds; the results reported in Sections 4 and 5 are the average of scores from four models. See Table 5 in Appendix for the hyperparameters and detailed settings.

2.3 Evaluation

Dataset: We used BLiMP (Warstadt et al., 2020), a benchmark of English grammatical judgment test to evaluate the models' L2 linguistic generalization. The dataset consists of 12 test suites; each corresponds to a specific linguistic phenomenon and falls into one of four coarse linguistic categories: morphology, syntax, semantics, and syntax&semantics. Each test suite has 1,000 minimal sentence pairs. Each pair consists of grammatically acceptable and unacceptable ones as follows:

(1) a. Many teenagers were helping themselves.

b. * Many teenagers were helping **herself**.

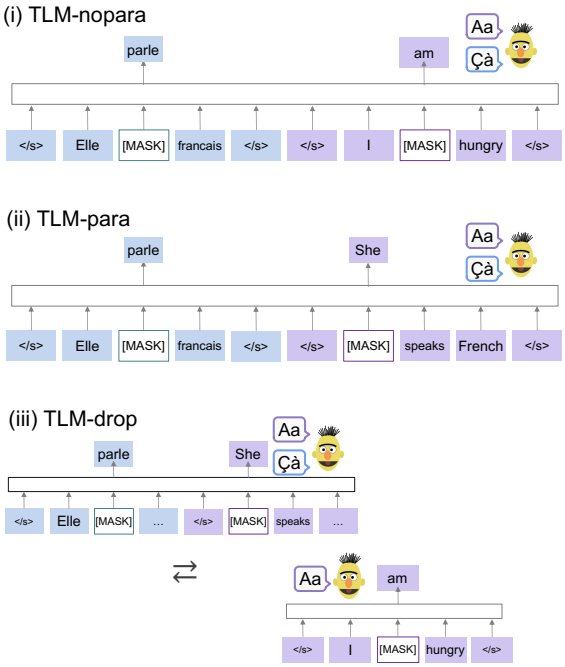

Figure 2: Training settings investigated in Section 3: (i) L1–L2 text pairs without translation relationship (TLM-nopara), (ii) translation pairs (TLM-para), and (iii) a mixed setting where parallel L1 text is removed every other epoch (TLM-drop).

Grammatical judgement: To select one sentence in each pair, we adopted pseudo-perplexity, commonly used in exploring the linguistic behaviors of LMs (Lau et al., 2020). Specifically, if the model can assign a lower pseudo-perplexity to the grammatical sentence than to the paired ungrammatical one, we regard it as correct. Following Salazar et al. (2020), psuedo-perplexity (PPPL) of sentence $\boldsymbol{s} = [w_1, w_2, \cdots, w_n]$ is computed using the bidirectional LM θ:

$$\mathrm{PPPL}(\boldsymbol{s}) = \prod_{t=1}^{n} p_\theta(w_t | \boldsymbol{s}_{\setminus w_t})^{\frac{1}{|\boldsymbol{s}|}} \quad , \qquad (1)$$

where w_t denotes the t-th token in sentence, and $\boldsymbol{s}_{\setminus w_t}$ denotes all the tokens in the sentence except for w_t; $[w_1, \cdots, w_{t-1}, w_{t+1}, \cdots, w_n]$. The probability of w_t given its bidirectional context $\boldsymbol{s}_{\setminus w_t}$ is calculated by the model θ. Based on the selected sentences, we calculated an accuracy score on each test suit of BliMP. We also report the macro-average of accuracies among all the test suits. Note that all the accuracy scores reported in the tables/figures in this paper are multiplied by 100 for readability.

3 Preliminary experiment: L2 exposure configurations

First, we investigate the inductive bias of L2 training settings. While existing studies use *parallel* data as an input for cross-lingual training (Conneau and Lample, 2019), we investigate the bias in this setting from L2 grammar acquisition perspectives.

Settings: We set up the three training settings with different input data: (i) L1–L2 text pairs without the translation relationship (TLM-nopara), (ii) L1-L2 translation pairs (TLM-para), and (iii) a mixed setting where L2 text concatenated with L1 parallel text is used as input or only L2 text is used as input (TLM-drop)[5]. An overview of the experimental settings is shown in Figure 2 (see details in Appendix A). Note that the original XLM (Conneau and Lample, 2019) adopts a setting similar to the TLM-drop. In this experiment, we report the macro-average BLiMP accuracies across the test suites. Table 2 shows the results (see Table 7 for the results in fine-grained test suits).

Translation pairs does not facilitate L2 acquisition: One notable point in Table 2 is that the results in the TLM-nopara setting were better than those in the TLM-para setting.[6] This suggests that parallel data input does not facilitate L2 acquisition. Perhaps, the TLM-para task was too easy for LMs to learn syntactic knowledge; the TLM-para task could partially be solved solely by relying on lexical knowledge, i.e., capturing the lexical correspondences between the tokens in L1 and L2 sentences and predicting the word found only in one of them. In this sense, the TLM-nopara setting, by contrast, might impose a more difficult problem on LMs and promote their effective learning of linguistic knowledge.

Switching between using L2 text with and without its parallel L1 text during training facilitates L2 acquisition: Another notable point is that the TLM-drop was the most effective for acquiring linguistic knowledge in L2 for LMs.[6] Since we switch between using L1 text as input or not every epoch, there is a possibility that monolingual and bilingual training have a complementary positive effect. In addition, this might mitigate the training-inference mismatch in evaluating LMs' linguistic knowledge using BLiMP. Here, a single sentence is used as input, which is compatible with the phase of using only L2 text during the TLM-drop training. In subsequent experiments, we will use the TLM-drop setting as it was the most effective training setting for L2 grammar acquisition.

Model (TLM)	Settings		First language			
	para.	*drop*	Fr	De	Ru	Ja
nopara			52.0	57.6	51.2	52.5
para	✓		51.1	53.6	48.9	51.3
drop	✓	✓	**58.0**	**61.1**	**52.8**	**56.2**

Table 2: Performance of bilingual LMs on BLiMP in different training settings. The *para.* column indicates whether parallel corpus was used. The *drop* column indicates whether the L2-side input is removed every other epoch.

4 Experiments: L1→L2 effects on linguistic generalization

We investigate how *pretraining with L1* affects L2 grammar acquisition in LMs. We exploratorily compare the linguistic generalization of LM trained in the settings with or without L1 pretraining. Table 3 shows the grammaticality judgment ability after additional training. The OVERALL column indicates the macro-average accuracy score across the grammar items. The Δ rows show the difference in BLiMP accuracy between models with and without pretraining. Here, the models without pretraining were trained only with bilingual corpus without L1 monolingual corpus pretraining.

4.1 L1s promote L2 generalization

Table 3 shows the effect of pretraining with L1 on L2 grammar acquisition. Most of the Δ values are positive; i.e., the models pretrained with L1s achieved better results than those without pretraining. This demonstrates that pretraining in a particular language generally improves English grammatical ability.[7] This positive effect is in light of the assumptions that different languages share some grammatical universals, and learners could use such a common property in language trans-

[5] In other words, the bilingual and monolingual settings are switched alternately for each epoch.

[6] Statistical differences between the settings are tested across seeds×languages with Mann-Whitney U tests (p=4.6e-2 for TLM-para and TLM-nopara, p=1.0e-2 for TLM-nopara and TLM-drop)

[7] The standard deviation over the different seeds is an average of 1.2 (0–100 scale) for the scores in Table 3, which means that the seed randomness does not overturn the conclusion.

			Morphology				Syntax				Semantics		Syntax & Sem.	
Lang.	L1	OVERALL	ANA. AGR	D-N AGR	IRREGULAR	S-V AGR	ARG. STR	ELLIPSIS	FILLER-GAP	ISLAND	NPI	QUANTIFIERS	BINDING	CTRL. RAIS.
Fr	✓	58.0	55.8	69.5	73.0	60.4	55.4	67.7	54.6	52.2	40.5	56.5	51.8	58.6
	Δ	5.3	2.3	14.5	3.8	8.8	7.2	17.0	4.5	−0.8	3.5	−1.2	1.9	1.5
De	✓	61.1	43.1	68.7	69.3	67.0	53.1	63.5	68.2	47.7	54.6	80.5	65.2	52.2
	Δ	5.2	5.9	11.1	−11.5	14.3	4.6	14.7	4.8	4.5	9.8	4.6	1.4	−2.2
Ru	✓	52.8	52.9	58.6	72.7	54.9	47.0	54.2	52.4	49.3	32.8	56.2	40.7	61.4
	Δ	0.7	−3.1	3.2	0.3	2.9	0.5	5.2	4.1	−1.9	1.1	−4.5	−0.1	0.3
Ja	✓	56.2	61.5	65.8	70.5	53.0	52.1	55.3	51.3	54.0	41.0	50.6	61.0	57.8
	Δ	1.5	0.7	5.0	−3.3	1.1	2.0	4.3	1.1	4.6	2.2	−2.9	0.8	2.3

Table 3: English (L2) grammatical knowledge of bilingual LMs with different L1s. OVERALL indicates the macro-average accuracy over all linguistic phenomena. The rows with ✓in the L1 column exhibit the accuracy of bilingual LMs on BLiMP. The rows with Δ in the L1 column exhibit the performance difference between the LMs with and without L1 pretraining. The coarse categories (e.g., Morphology) are from the metadata of BLiMP.

L1	Morph.	Syntax	Semantics	Syn.&Sem.
Fr	7.3	7.0	1.2	1.7
De	5.0	7.2	7.2	−0.4
Ru	0.8	1.9	−1.7	0.1
Ja	0.9	3.0	−0.3	1.5
Avg.	**3.5**	**4.8**	**1.6**	**0.7**

Table 4: Performance difference between the LMs with and without L1 pretraining for each coarse category of grammatical items (morphology, syntax, semantics, and syntax&semantics).

fer (Cook, 1985). Ri and Tsuruoka (2022) also reported that pretraining in a natural language other than English improves the overall syntactic parsing performance in English. Our results are consistent with such positive effects. Besides, our experiment provides the results of each of more fine-grained grammatical items related to morphological, syntactic, and semantic phenomena.

4.2 Differences in L1s

The Δ values in the OVERALL column in Table 3 differ across the L1s. French is the highest, followed closely by German, and Japanese and Russian are far behind these two languages; pretraining in French and German is much more effective than in Japanese and Russian. This ordering shows parallels with the presumed language learning difficulty order: French, German, Russian, and Japanese. This suggests that the difficulty of acquiring an L2 grammatical ability is somewhat close between LMs and humans.

4.3 Differences in grammar items

The Δ scores in Table 3 exhibit that different grammatical items obtain different degrees of gains. Table 4 shows the average Δ scores for each course grammar category. There was a general tendency for morphological and syntactic items to get larger gains from the L1 pretraining than semantic and syntax&semantic items except for particular settings, e.g., IRREGULAR. It has been shown that linguistic phenomena related to semantics such as NPI (negative polarity item) and QUANTIFIERS were relatively difficult for LMs to learn (Warstadt et al., 2020). Based on this, there is a concern that LMs failed to learn enough such linguistic knowledge in L1 to transfer it to another language.

4.4 Differences in L1×grammar-item

Notably, in specific combinations of L1s and grammar items, the L1 pretraining hurt the L2 generalization, i.e., negative transfer problem. For example, the performance in the IRREGULAR item was not so much enhanced or even degraded by L1 pretraining. The IRREGULAR (Irregular forms) item targets English-specific irregular verb conjugations; its less effect by L1 pretraining is due to the concern that the *irregular* patterns generally could

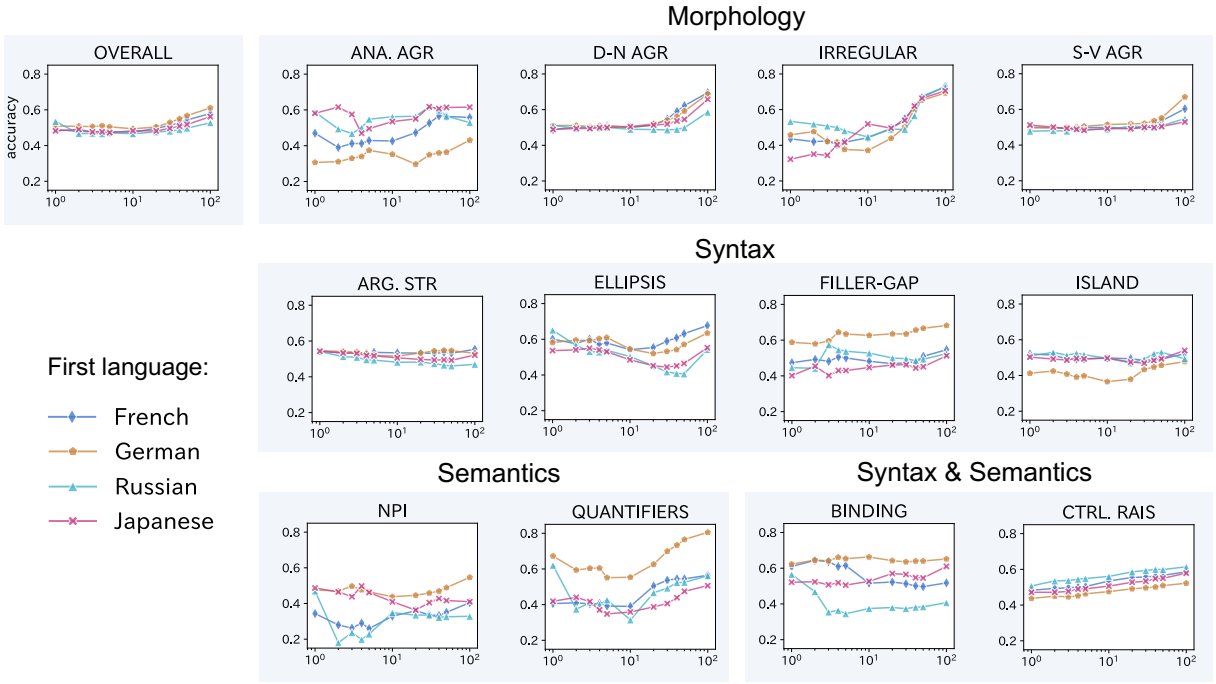

Figure 3: Grammar learning trajectories in each test suite on BLiMP (the L2 side). The x-axis denotes the epoch during L2 acquisition, and the y-axis denotes the accuracy in the corresponding test suite.

not be predictable by other language's knowledge.

We also found that the same grammatical item was affected differently depending on the L1. For example, the Δ values on the FILLER-GAP item in Table 3 differ across the L1s, e.g., 4.8 in German and 1.1 in Japanese. At least in this FILLER-GAP aspect, there is an interesting parallel between our results and linguistic notions; Japanese is the only language where gap precedes filler in wh-construction among the L1s we used, and the transfer from Japanese to English was indeed limited ($\Delta = 1.1$) compared to other L1s. There might be a possibility that such linguistic (dis)similarities are reflected in the results. Nevertheless, concluding the exact consistency between our L1$\times$grammar-item results (Table 3) and the L1–L2 grammatical similarity requires further interdisciplinary research.

5 Analysis: acquisition process

This section sheds light on the *process* of L2 acquisition. We investigate how L2 knowledge acquisition progresses (Section 5.1) and how original L1 knowledge changes during L2 acquisition (Section 5.2). As for the L1 knowledge during L2 acquisition, there is a concern, for example, that LMs exhibit catastrophic forgetting about their L1.

5.1 L1→L2 effects

Settings: We evaluate the L2 linguistic knowledge of intermediate checkpoints of LMs (Figure 3). Specifically, we evaluate LMs after {1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 100} training epochs. Note that we analyzed the same models used in Section 4; the results after 100 epochs are the same ones reported in Section 4.

General improvement after dozens of epochs: The trajectory of the OVERALL scores in Figure 3 suggest that linguistic ability generally improves along with the number of epochs. There was a tendency for large improvements to emerge after dozens of epochs; in other words, the models began to acquire L2 knowledge after seeing the same examples many times, e.g., 50-100 times. Note that humans are argued to acquire a vocabulary after encountering the same word about 12 times (Nation, 2014), and of course, the lexical and syntactic acquisition is not comparable, but the observation that the L2 knowledge improves after 50-100 rounds of the corpus may be in the direction that LMs are inefficient at acquiring a new language.

Differences in grammar items: Focusing on the general trajectory shapes for each grammatical item, we observed at least four patterns: (i) spike-at-the-end (D-N AGR., IRREGULAR, S-V

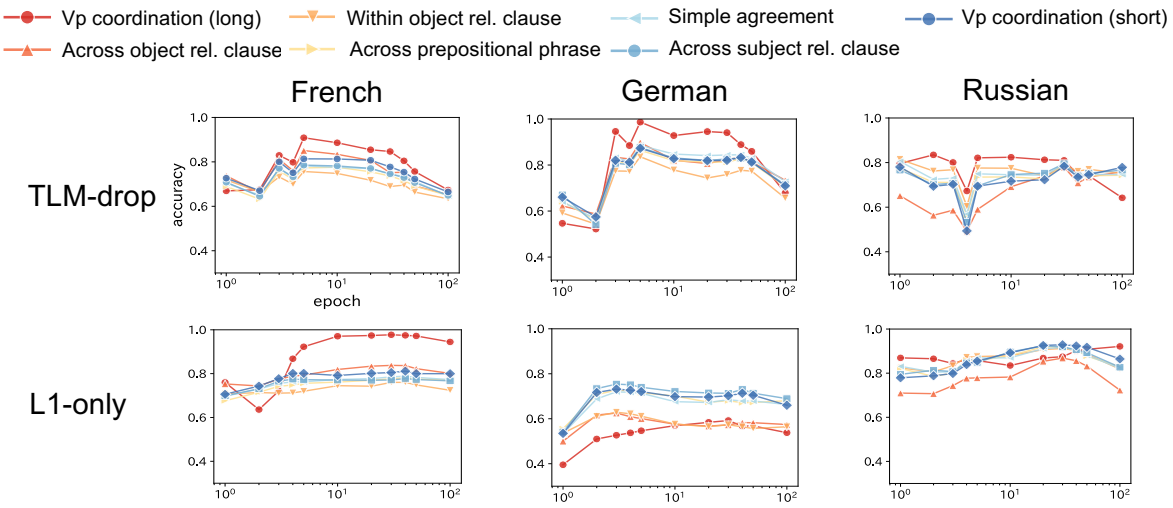

Figure 4: Grammar learning trajectories in each test suite on CLAM (the L1 side). The x-axis denotes the epoch during L2 acquisition, and the y-axis denotes the accuracy in the corresponding test suite. The upper parts show the scores of the bilingual LMs (L1 pretraining and bilingual training). The lower parts show the scores of the L1-only LMs (L1 pretraining and further training on L1 texts collected from the parallel corpus).

AGR.), (ii) flat (ARG.STR., CTRL.RAIS., ISLAND), (iii) bumpy (ANA.AGR., ELLIPSIS, NPI, QUANTIFIERS), and (iv) mixed (FILLER-GAP, BINDING). In addition, these groups roughly mirror the linguistic categories of the grammar items (morphology, syntax, semantics, and syntax&semantics); for example, all the items in the spike-at-the-end group are morphological phenomena, while all the semantic categories (NPI, QUANTIFIERS) yielded the bumpy patterns. Note that existing studies reported that low-level (e.g., morphological) linguistic skills could be acquired earlier and vice versa (Liu et al., 2021; Blevins et al., 2022); but at least in our cognitively-inspired bilingual training scenario, we did not observe such an explicit tendency.

Inter-L1s differences: In terms of the change of inter-L1s differences of accuracies in each grammar item, there are several different patterns: (i) converging (IRREGULAR, ISLAND), (ii) diverging (ARG.STR., BINDING, D-N AGR., S-V AGR.), and (iii) none of them. Considering IRREGULAR as an example of the converging group, the accuracies were substantially different across the L1s in the initial stage of training; these differences, however, gradually reduced along with epochs. On the other hand, considering S-V AGR. as an example of the diverging group, the accuracies gradually differed in the latter stage of training among the LMs. In the third group, the inter-L1s accuracy differences remain the same or unstable during L2 tranining (ANA.AGR., CTRL.RAIS., ELLIPSIS, FILLER-GAP, NPI, QUANTIFIERS). At least the former two groups imply that pretraining with different L1s differently affects the efficiency of L2 acquisition (e.g., different slopes for different L1s).

To sum up, we clarified that different L1s and grammar items exhibit different learning dynamics of LMs. The cognitive plausibility of these patterns could be the next important investigation.

5.2 L2→L1 effects

In contrast to the previous analysis, which analyzed the impact in the L1→L2 direction, we further analyze the L2→L1 impact. In applied linguistics, L1–L2 impact in both directions is of interest; for example, it is reported that the L1 ability is sometimes hurt by the increase of L2 exposure (Kecskes, 2008; Haman et al., 2017).

Settings: In the same way as Section 5.1, we evaluate the grammatical knowledge of LMs during L2 training, but the focus is on the **L1** knowledge. We used a multilingual benchmark of grammatical judgment test CLAMS (Mueller et al., 2020). This dataset consists of seven syntactic test suites for several languages. Similarly to BLiMP, the dataset consists of pairs of sentences, where one is grammatical, and the other is ungrammatical. We report the accuracy scores in terms of whether the LMs could assign lower pseudo-perplexity to the grammatically correct sentence. We analyze French-L1,

German-L1, and Russian-L1 LMs since this dataset covers these three L1s.

As a baseline, we also evaluate the L1-only LMs that are first pretrained in L1, then additionally trained with the L1-side texts collected from the corresponding parallel corpus; that is, the only difference between bilingual LMs and L1-only baselines was the presence of L2 texts during the L2 acquisition phase.

L2 effects once occur, but diminish: Figure 4 shows the results (see Table 8 for the exact scores). We found that the L1 knowledge is greatly influenced by the L2, especially at the initial stage of L2 training. For example, the French-L1 and German-L1 LMs gained positive effects, and the Russian-L1 LM got a negative influence (the top row of Figure 4). In addition, in the latter stage, the L2 effects on the L1 knowledge gradually fade, in either a good or bad sense, and the L1 linguistic knowledge is converging to the original level. For example, French-L1 and German-L1 LMs exhibited better performance after bilingual language modeling once, e.g., at the 10 epochs, but the gain decreased after further bilingual training.

L2 negatively affects L1 knowledge: In Figure 4), the L1 knowledge in bilingual LMs was competitive or even inferior to that in L1-only LMs in the end, although there were also exceptional cases in specific grammar items in German. For example, in the French-L1 LMs, the L1 syntactic generalization performance after L2 training converged below 0.7 points, while the L1-only baseline model generally achieved above 0.7 points. Contrasting the L1 results with L2 ones (Section 5.1), there is an asymmetry effect between L1 and L2. From a perspective of developing linguistically better multi-lingual LMs, these suggest that balancing the linguistic knowledge of the languages, especially enhancing the L1 knowledge, during language transfer is challenging even if the model is exposed to L1 during bilingual training. Addressing these challenges with, for example, some regularization will be a promising direction from an engineering perspective.

6 Related work

Cognitively-motivated analysis of neural LMs: Investigations into the ability of neural models in language acquisition began in the 1980s with the interest of whether language can be acquired without innate knowledge (Rumelhart and McClelland, 1986; Pinker and Prince, 1988). The initial investigation was made with simple neural networks; after the development of Neural NLP (Manning, 2015), the classical questions posed by cognitive science are currently revisited (Kirov and Cotterell, 2018). A typical movement is a growing interest in the *probing* of neural LMs' linguistic knowledge (Linzen et al., 2016; Warstadt and Bowman, 2020). In this context, our study analyzes L2 acquisition in neural LMs, while existing studies have typically focused on L1 acquisition.

Language transfer in computational models: Language transfer of NLP models is actively researched from both engineering and scientific perspectives. In the engineering context, to mitigate the English-centric focus of NLP techniques, models that can handle more languages have been developed (Dong et al., 2015; Conneau and Lample, 2019; Conneau et al., 2020). From the scientific perspective, the mechanism and linguistic properties of LMs' language transfer have been explored (Pérez-Mayos et al., 2021; Tyler et al., 2022; Blevins et al., 2022), sometimes beyond the transfer between natural languages (Ri and Tsuruoka, 2022; Papadimitriou and Jurafsky, 2020). One of the motivations of such analyses is to quantify the transferable universals behind (non-)languages. Notably, simulation of L2 acquisition is also explored from pedagogical motivations (Settles et al., 2018).

Language transfer in humans: L2 acquisition/learning has long been studied in applied linguistics, psycholinguistics, and pedagogy fields (Krashen, 1981; Hatch, 1983; Ellis, 2010). These fields articulated several hypotheses/theories on human language learning, e.g., input hypothesis (Krashen, 1977). Analyzing the LMs' L2 acquisition in a more direct light with these hypotheses would be interesting for future work.

7 Conclusions

We have investigated the L2 acquisition of LMs, especially focusing on their grammatical knowledge in L1 and L2. Specifically, we have trained bilingual LMs under a similar scenario to the human L2 acquisition and then analyzed their cross-lingual transfer. Our experiments have demonstrated that L1 pretraining promotes their linguistic generalization in L2, and there are interesting variations in L1 pretraining effects with respect to the L1 choice,

training settings, and grammar items. The results have also implied that their L2 acquisition is not human-like in particular aspects.

Limitations

Coverage of experiments

There is room for further exploration in terms of model architectures, data, and evaluation settings in our study. Experiments in more diverse settings would enhance the generality of the conclusions.

Models: The architecture was fixed to XLM (14M), although we tested four models with different seeds in each setting. Specifically, testing unidirectional LMs will be in a reasonable direction, considering the humans' incremental language processing. Related to this, the measure of pseudo-perplexity might also induce unintended biases in our results, this is a common metric in NLP though. In addition, there are different methods to fine-tune the model to multiple languages, e.g., using adapters. Comparing these methods with our scheme will be an interesting direction.

Data: There are possible variations of L1–L2 combinations. Although we selected L1 from one of the four languages (German, French, Russian, Japanese) and fixed L2 as English, increasing the coverage of languages will lead to more generalized conclusions. Furthermore, the performance of LMs was generally not so good on the BLiMP dataset. We suspect that this is due to the limited L2 training data size; it is also worth exploring scaling up the experiments into typical NLP experiments.

Evaluation: While our focus is on morphology, syntax, and semantic generalization, L2 acquisition studies are conducted from broader perspectives, such as the growth of vocabulary size. In addition, our observations are from the perspective of the LMs; the contrast between LMs' and humans' L2 learning is more important from an interdisciplinary perspective.

Performance was overall poor

One reviewer is concerned that our results on the BLiMP are generally near the chance level, and it may be difficult to derive findings from such poor results. We thank the reviewer and would like to share our thoughts and limitations here.

First, comparisons to the chance rate are not always meaningful. In BLiMP, the task is typically to select a correct generalization over an incorrect one. Occasionally, neural models overly prefer incorrect generalizations more than chance level. For instance, in the linear vs. hierarchical generalization contrast, neural models often favor linear, causing accuracy to drop near 0, far below chance (McCoy et al., 2018). In such cases, achieving accuracies around 50 indicates more than random guessing, as models avoid an excessive preference for incorrect generalizations, moving toward a more neutral stance. Thus, we believe that it is also worthwhile to observe how much performance improves from below the chance level.

Furthermore, in BLiMP's finer-grained test suits, our models sometimes exhibit an accuracy of 0 or 100 (resulting in an overall score of around 50.0), highlighting that our models do not always act as random guessing baselines. The full results across more fine-grained test suits are shown in Appendix 7.

Ethics Statement

There might be a possibility that the texts we used (CC-100 and Tatoeba) have socially biased, despite their popular use in the NLP community. We adopted cognitively-plausible restricted settings with respect to data size, which can potentially be aligned with environmentally friendly, green NLP.

Acknowledgements

We would like to express our gratitude for the anonymous reviewers who provided many insightful comments that have improved our paper. Special thanks also go to the members of NAIST and Tohoku NLP Laboratory for the interesting comments and energetic discussions. This work was supported by JSPS KAKENHI Grant Number JP19K20351.

References

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of EMNLP*, pages 3575–3590, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Joan L. Bybee. 2013. 49 Usage-based Theory and Exemplar Representations of Constructions. In *The Oxford Handbook of Construction Grammar*. Oxford University Press.

Barry Chiswick and Paul Miller. 2004. Linguistic distance: A quantitative measure of the distance between english and other languages. Technical Report 1246, Institute of Labor Economics (IZA).

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Vivian J Cook. 1985. Chomsky's universal grammar and second language learning. *Applied linguistics*, 6(1):2–18.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of NAACL*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of ACL*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Rod Ellis. 2010. Second language acquisition, teacher education and language pedagogy. *Language teaching*, 43(2):182–201.

Barbara F Grimes and JE Grimes. 2002. Ethnologue: Languages of the world. 14, h edition. *Dallas, TX: SIL International*.

Ewa Haman, Zofia Wodniecka, Marta Marecka, Jakub Szewczyk, Marta Białecka-Pikul, Agnieszka Otwinowska, Karolina Mieszkowska, Magdalena Łuniewska, Joanna Kołak, Aneta Miękisz, Agnieszka Kacprzak, Natalia Banasik, and Małgorzata Foryś-Nogala. 2017. How does L1 and L2 exposure impact L1 performance in bilingual children? evidence from Polish-English migrants to the united kingdom. *Front. Psychol.*, 8:1444.

Evelyn Marcussen Hatch. 1983. *Psycholinguistics: A second language perspective.* ERIC.

Istvan Kecskes. 2008. The effect of the second language on the first language. *Babylonia*, 2(2):31–34.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Philipp Koehn, 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, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Stephen Krashen. 1977. Some issues relating to the monitor model. *On Tesol*, 77(144-158).

Stephen Krashen. 1981. Second language acquisition. *Second Language Learning*, 3(7):19–39.

Stephen D Krashen, Michael A Long, and Robin C Scarcella. 1979. Age, rate and eventual attainment in second language acquisition. *TESOL Quarterly*, 13(4):573–582.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher D. Manning. 2015. Last words: Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *In Proceedings of CogSci*, pages 2093–2098, Madison, WI.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of ACL*, pages 5523–5539, Online. Association for Computational Linguistics.

Paul Nation. 2014. How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, 26(2):2.

Graham Neubig and Shinsuke Mori. 2010. Word-based partial annotation for efficient corpus construction. In *Proceedings of LREC*, Valletta, Malta. European Language Resources Association (ELRA).

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of ACL*, pages 529–533, Portland, Oregon, USA. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of EMNLP*, pages 6829–6839, Online. Association for Computational Linguistics.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1):73–193.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of ACL*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

David E Rumelhart and James L McClelland. 1986. On learning the past tenses of english verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 216–271. MIT Press, Cambridge, MA.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of ACL*, pages 2699–2712, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of LREC*, Istanbul, Turkey. European Language Resources Association (ELRA).

Chang Tyler, Tu Zhuowen, and Benjamin Berge. 2022. The geometry of multilingual language model representations. In *Proceedings of EMNLP*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? In *In Proceedings of CogSci*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Guillaume Wenzek. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of LREC*, pages 4003–4012, Marseille, France. European Language Resources Association.

dropout	0.1
attention_dropout	0.1
accumulate_gradients	4
emb_dim	256
ffn_embed_dim	1024
gelu_activation	True
Optimizer	adam_inverse_sqrt
	lr=0.00020
	eps=0.000001
	warmup_updates=30000
	beta1=0.9
	beta2=0.999
	weight_decay=0.01
epoch	100
n_heads	8
n_layers	12
clip_grad_norm	1.0
amp	2
fp16	True

Table 5: Hyperparameters of the LMs

Name	Version	License
fastBPE	0.1.0	MIT License
kytea	0.4.7	Apache 2.0
mosesdecoder	0.4.0	LGPL 2.1
XLM	0.1.0	CC BY-NC 4.0
BLiMP	0.1.0	CC BY-NC 4.0
CC100 (CC-Net)	1.0.0	MIT License
CLAMS	0.1.0	Apache 2.0
Tatoeba	v2022-03-03	CC–BY 2.0 FR

Table 6: The versions and licenses of used tools and datasets. These tools and datasets used in this study were designed for the purposes of research and language learning.

A Experimental Procedure

We list the hyperparameters in Table 5. The versions and licenses of used tools and datasets are listed in Table 6.

L1 Acquisition: We used mosesdecoder (Koehn et al., 2007) as French, German and Russian tokenizer, kytea[8](Neubig et al., 2011; Neubig and Mori, 2010) as Japanese tokenizer and segmented words into subwords with fastBPE[9](Sennrich et al., 2016) The dataset was split into train/dev/test in the ratio of 8:1:1. We set 14,000 vocabulary size for any language. We trained our models with 4 parallel GPUs (VRAM 48G), which took 6 days per model.

L2 Acquisition: We added English tokens from the parallel corpus into BPE codes and vocabulary used in L1 Acquisition and removed duplicated tokens and vocabulary. As for models not using a monolingual corpus, we created the codes and vocabulary using a parallel corpus of both L1 and L2. We used mosesdecoder (Koehn et al., 2007) as English tokenizer. As for other languages, we use tokenizers the same as L1 Acquisition. The dataset was split into train/dev/test in the ratio of 8:1:1. As we increased the number of vocabularies in the embedding layer, the weights/biases in the final layer were also increased. Our four LMs were trained with different 3 seeds and reported their averages as results. Compared models in our preliminary experiment (Sec. 3) are shown in Figure 2. We trained our models with 2–4 GPUs (VRAM 48G), which took around 5 hours per model.

[8] http://www.phontron.com/kytea/
[9] https://github.com/glample/fastBPE

Coarse	Specific	Challenge	First Language			
			Fr	De	Ru	Ja
Morphology	ANA.AGR	anaphor_gender_agreement	61.5	23.7	52.7	57.7
		anaphor_number_agreement	50.1	62.5	53.1	65.4
	D-N AGR	determiner_noun_agreement_1	80.8	74.5	69.1	67.9
		determiner_noun_agreement_2	70.1	76.5	62.6	79.8
		determiner_noun_agreement_irregular_1	65.8	63.0	58.3	54.5
		determiner_noun_agreement_irregular_2	64.9	72.3	56.5	78.3
		determiner_noun_agreement_with_adj_1	77.9	69.2	59.9	61.1
		determiner_noun_agreement_with_adj_2	62.8	66.9	56.1	65.7
		determiner_noun_agreement_with_adj_irregular_1	72.7	64.1	54.0	50.5
		determiner_noun_agreement_with_adj_irregular_2	61.1	63.0	51.9	68.4
	IRREGULAR	irregular_past_participle_adjectives	75.1	61.6	95.3	79.1
		irregular_past_participle_verbs	70.9	77.0	50.0	62.0
	S-V AGR	distractor_agreement_relational_noun	50.9	65.5	39.8	42.6
		distractor_agreement_relative_clause	48.6	51.7	46.1	45.4
		irregular_plural_subject_verb_agreement_1	64.2	68.9	56.7	54.4
		irregular_plural_subject_verb_agreement_2	65.5	74.5	63.6	58.9
		regular_plural_subject_verb_agreement_1	65.9	72.2	60.7	59.5
		regular_plural_subject_verb_agreement_2	67.3	69.2	62.5	57.2
Syntax	ARG.STR	animate_subject_passive	65.1	55.9	51.1	53.9
		animate_subject_trans	44.7	44.1	31.7	37.3
		causative	39.6	53.4	35.8	38.8
		drop_argument	57.5	44.6	44.0	54.1
		inchoative	51.6	43.5	37.7	45.2
		intransitive	56.0	47.1	40.5	52.4
		passive_1	61.4	62.6	60.9	67.7
		passive_2	62.7	70.0	65.1	66.2
		transitive	59.8	56.8	56.1	53.8
	ELLIPSIS	ellipsis_n_bar_1	52.1	50.2	46.4	46.0
		ellipsis_n_bar_2	83.2	76.8	62.1	64.5
	FILLER-GAP	wh_questions_object_gap	34.7	87.2	36.6	31.4
		wh_questions_subject_gap	66.8	89.2	64.9	61.8
		wh_questions_subject_gap_long_distance	68.8	95.5	63.8	62.4
		wh_vs_that_no_gap	68.4	97.1	60.5	63.2
		wh_vs_that_no_gap_long_distance	61.3	97.8	56.1	53.5
		wh_vs_that_with_gap	39.8	6.6	40.1	38.8
		wh_vs_that_with_gap_long_distance	42.6	4.1	45.2	47.7
	ISLAND	adjunct_island	49.0	52.5	50.0	56.8
		complex_NP_island	47.7	43, 0	43.9	53.9
		coordinate_structure_constraint_complex_left_branch	33.7	40.2	41.1	39.0
		coordinate_structure_constraint_object_extraction	68.8	78.9	57.7	53.9
		left_branch_island_echo_question	36.6	37.2	27.1	37.2
		left_branch_island_simple_question	63.1	71.7	68.0	69.1
		sentential_subject_island	50.5	50.1	52.3	55.4
		wh_island	68.0	8.1	54.5	66.6
Semantics	NPI	matrix_question_npi_licensor_present	46.2	46.0	14.3	30.9
		npi_present_1	52.1	94.5	41.3	72.3
		npi_present_2	56.4	94.8	39.5	72.5
		only_npi_licensor_present	1.0	0.2	0.1	0.2
		only_npi_scope	35.9	55.2	50.4	52.1
		sentential_negation_npi_licensor_present	32.1	34.2	39.6	11.8
		sentential_negation_npi_scope	59.6	57.6	44.6	47.1
	QUANTIFIERS	existential_there_quantifiers_1	85.2	84.0	69.6	62.3
		existential_there_quantifiers_2	7.5	61.2	4.2	2.8
		superlative_quantifiers_1	45.2	92.0	60.6	55.4
		superlative_quantifiers_2	87.9	84.9	90.3	82.0
Syntax & Semantics	BINDING	principle_A_c_command	61.4	49.6	69.8	53.5
		principle_A_case_1	52.0	99.8	11.4	99.9
		principle_A_case_2	62.0	58.7	54.8	49.6
		principle_A_domain_1	41.3	93.7	0.7	83.9
		principle_A_domain_2	51.3	62.9	49.9	45.2
		principle_A_domain_3	52.1	49.4	48.5	48.7
		principle_A_reconstruction	42.4	42.2	49.6	46.3
	CTRL. RAIS	existential_there_object_raising	68.2	53.8	68.5	69.9
		existential_there_subject_raising	55.1	53.3	69.7	51.1
		expletive_it_object_raising	70.2	54.7	67.9	69.1
		tough_vs_raising_1	56.4	44.6	27.2	62.3
		tough_vs_raising_2	43.2	54.8	73.8	36.8

Table 7: Results for each fine-grained test suit in BLiMP.

Model	L1	Epoch	long_vp_coord	obj_rel_within_anim	simple_agrmt	vp_coord	obj_rel_across_anim	prep_anim	subj_rel
TLM-drop	Fr	5	90.9	75.7	78.2	81.4	85.1	77.3	78.5
		50	75.7	66.4	71.2	72.3	69.2	69.1	70.7
		100	67.3	63.4	64.6	66.5	65.2	64.9	64.9
	De	5	98.7	83.6	88.5	87.3	89.8	86.4	87.6
		50	85.9	77.4	82.9	81.3	81.3	81.0	81.4
		100	67.9	65.6	72.8	71.0	73.4	70.6	71.4
	Ru	5	82.1	77.6	75.0	69.4	59.0	73.5	69.6
		50	74.1	77.0	74.7	74.6	73.5	74.1	75.3
		100	64.2	75.1	74.7	77.9	75.9	74.3	76.6
L1-only	Fr	5	92.3	79.3	72.0	75.6	76.3	77.3	80.1
		50	97.2	82.3	74.9	77.6	78.3	77.3	80.0
		100	94.5	80.1	72.4	77.1	77.3	76.7	80.9
	De	5	54.7	60.1	61.1	71.6	71.3	74.0	72.1
		50	56.8	58.2	55.7	67.2	67.9	71.4	70.5
		100	53.8	57.4	56.4	68.0	66.3	68.9	66.0
	Ru	5	85.7	77.9	87.5	85.6	86.9	84.9	85.5
		50	90.8	83.2	88.6	87.7	88.1	89.3	91.8
		100	92.1	72.3	81.8	82.9	82.2	82.7	86.4

Table 8: Results for each fine-grained test suit in CLAMS.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Limitation section

☑ A2. Did you discuss any potential risks of your work?
In Ethics Statement section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract section is located before introduction section at page 1. The section number of introduction is 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In Section 2, 3, 4, 5 and Appendix

☑ B1. Did you cite the creators of artifacts you used?
In Appendix section

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Appendix section

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Section 2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In our Appendix

C ☑ Did you run computational experiments?

In Section 2, 3, 4, 5 and Appendix

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Section 2 and Appnedix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Appendix section

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Appendix section

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Appendix section

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.