Once a Tree, Always a Tree?
Cross-lingual Transfer of Hierarchical Generalization in Language Models

Shin Yamaguchi [1] **Masaya Taniguchi**[2]
Tatsuki Kuribayashi[3,4] **Taro Watanabe**[1] **Yoshinobu Kano**[5]

[1]NAIST [2]RIKEN [3]MBZUAI [4]Tohoku University [5]Shizuoka University
yamaguchi.shin.yw4@naist.ac.jp masaya.taniguchi@riken.jp
tatsuki.kuribayashi@mbzuai.ac.ae taro@is.naist.jp
kano@inf.shizuoka.ac.jp

Abstract

Language models (LMs) often exhibit successful cross-lingual transfer through continued (pre)training, still their underlying process remains unclear. One possibility is that LMs can leverage the evidence toward language-universal principles, namely, hierarchical structure, from a source language (L1) to a new language (L2), while many other non-linguistic explanations, e.g., improved initialization of position embeddings, can be potentially possible. In this study, to quantify the cross-linguistic transferability of linguistic evidence in cross-lingual LM training, we introduce a new experimental design, namely, a cross-lingual poverty-of-stimulus (CPOS) setting that ablates how LMs' hierarchical generalization in L2 can be led by structural evidence provided in L1. We contrast counterfactual L1 and L2 learning scenarios where the L2 corpus is ambiguous between linear and hierarchical grammar rules, while the amount of positive evidence for the hierarchical rule is controlled on the L1 side. Our experimental results show that LMs can partially, yet not perfectly, transfer L1 structural evidence to L2, yielding only a small but detectable hierarchical bias, while the strength of transfer depends on language pairs. We further ablate the conditions for the successful transferring, motivated by recent reports on the substantial impact of vocabulary overlap. We indeed find that vocabulary overlap is crucial even for the cross-lingual transfer of hierarchical syntactic generalization. This suggests That this also explains the language-pair-dependent tendencies in the main experiments and suggests that the transfer of linguistic evidence in continued new-language learning is only conditionally achievable.

1 Introduction

Cross-lingual transfer of language models (LMs) generally succeeds in learning a new language and, in practice, helps propagate LMs developed in one language to others (Conneau et al., 2020). Even from a scientific perspective, cross-lingual transfer has served as a window into whether and how LMs encode language-universal properties from source data (L1) and use them for learning a new language (L2) (Papadimitriou and Jurafsky, 2020; Ri and Tsuruoka, 2022). This even potentially bridges LMs into psycholinguistics as a tool for characterizing human L2 learning, regarding LMs' cross-lingual transfer as a reference (Oba et al., 2023; Yadavalli et al., 2023; Aoyama and Schneider, 2024). Despite its ubiquity, the underlying mechanisms and biases behind LMs' cross-lingual transfer remain unclear. One prominent view is that

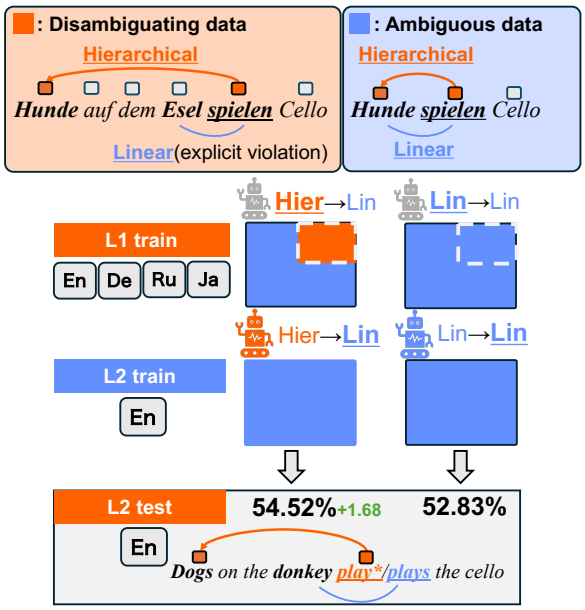

Figure 1: Overview of the experimental design. A language model is first pretrained on an L1 corpus that contains positive evidence for the hierarchical rule (**Disambiguating data**), and then continued-pretrained on an L2 corpus from which such evidence has been removed (**Ambiguous data**). Any above-chance hierarchical-syntax performance at L2 must therefore reflect cross-lingual transfer of the rule acquired during L1 training.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 60982–60998
October 24-29, 2026 ©2026 Association for Computational Linguistics

LMs identify and abstract language-universal properties from evidence provided in L1 data, especially the hierarchical organization of natural-language structure, and then leverage this evidence in L2 learning. Many other possibilities, however, can coexist alongside successful cross-lingual transfer, such as a mere scaling effect from increased pretraining data size or non-linguistic initialization artifacts, such as improved position embeddings learned from L1 data. Studying a causal effect from L1 to L2 transfer learning requires controlled, counterfactual ablations.

In this paper, we conduct targeted ablations of LMs' L1 and L2 transfer learning scenarios to isolate the transfer of language-universal, hierarchical properties that have long been studied in linguistics (Chomsky, 1980; Legate and Yang, 2002; Linzen et al., 2016; Gulordava et al., 2018). We specifically introduce a new experimental design, namely, a cross-lingual poverty-of-stimulus (CPOS) setting, in which the L2 corpus is ambiguous between linear and hierarchical rules of L2 grammar, and positive evidence for the hierarchical rule is available only on the L1 side. If LMs properly abstract the evidence in L1 to prefer hierarchical rules, increasing the amount of positive evidence on the L1 side will clearly improve the hierarchical generalization of L2 grammar. We implement such experiments using subject-verb number agreement constructions (see Figure 1) and control the number of direct evidence disambiguating the hierarchical agreement rule from a position-based one, as well as indirect evidence from deeply nested constructions implicitly clarifying the hierarchical structure[1].

We perform experiments with four L1 languages (i.e., English, German, Russian, and Japanese) to analyze language-dependent factors and evaluate LMs' generalization in the English language (L2). Our experiments demonstrate that exposure to positive evidence in L1 indeed enhances, though not perfectly, the hierarchical generalization in L2, confirming the cross-lingual transfer of positive evidence. We also observe some language-pair-dependent tendencies; for example, the German language as L1 leads to more successful generalization in L2 English than Japanese as L1. To clarify these results, we further analyze the lexical overlap effect, which has been reported as a dominant factor in predicting the success of cross-lingual transfer in downstream tasks (Kallini et al., 2025). Our analysis echoes the reported trends: eliminating vocabulary overlap, even using the same English corpora for both L1 and L2 but encoded with disjoint token IDs, leads to a clear failure of transfer. This suggests that the transfer of positive linguistic evidence toward the language-universal, hierarchical structure is only conditionally successful, and the primary condition is vocabulary overlap. That is, the vocabulary overlap is a fundamental path for successful cross-lingual transfer, even for the primitive language-generalization problem of syntactic rules, let alone downstream tasks (Kallini et al., 2025).

2 Backgrounds

2.1 Cross-lingual Transfer and the L2-LM Paradigm

Existing works examine cross-lingual transfer both as a practical recipe for extending LMs to new languages and as a scientific probe into the kinds of transferable knowledge LMs acquire from L1 data and how this biases their L2 learning and processing. This L2-LM paradigm (Aoyama and Schneider, 2024; Oba et al., 2023; Arnett et al., 2025; Yadavalli et al., 2023; Gao et al., 2025) organizes the latter perspective, casting continued pretraining from L1 to L2 as a bilingual setting: varying L1 while holding L2 fixed isolates the effect of the source language on L2 outcomes, and the resulting L2 trajectories provide a baseline for comparison against human L2 acquisition. We adopt this design and combine it with the linguistic and non-linguistic factors discussed below in a poverty-of-stimulus setting targeting hierarchical generalization (§2.2). Some studies show that typological distance between languages is related to the transfer effect (Yadavalli et al., 2023; Oba et al., 2023). To probe this transferability, prior work uses Test for Inductive Bias via Language Model Transfer (TILT)-based methods that test whether inductive biases acquired during source-language training transfer to target-language evaluations (Papadimitriou and Jurafsky, 2020; Aoyama and Schneider, 2024), as well as crosslinguistic priming effects across language pairs (Arnett et al., 2025).

A related line of work explicitly introduces inductive biases into LMs by pretraining them on non-natural inputs such as music or artificial languages (Papadimitriou and Jurafsky, 2020; Ri

[1]Code : https://github.com/shin0729/cpos; Data: https://huggingface.co/datasets/shin0729/cpos

and Tsuruoka, 2022), allowing controlled tests of whether source-language training induces a form of structural meta-learning for the target language. Other studies examine non-linguistic factors, including different conditions of language exposure (Constantinescu et al., 2025; Arnett et al., 2025) and parameter sharing across languages (Conneau et al., 2020).

Vocabulary Overlap in Cross-Lingual Transfer In addition to these factors, it has been shown that cross-lingual transfer in LMs depends on the degree of subword overlap between the paired-language vocabularies (Kallini et al., 2025; Škrjanec et al., 2026). In particular, Kallini et al. (2025) devise a controlled intervention in which the token IDs of the *native overlap* are selectively offset to enforce disjoint L1 and L2 vocabularies, and show that the presence of shared vocabulary drives cross-lingual transfer, improving downstream task performance. Škrjanec et al. (2026) likewise find that the cross-lingual alignment of embedding spaces is largely driven by vocabulary sharing. We adopt the same form of overlap control (§6).

2.2 Hierarchical Generalization and Poverty-of-Stimulus Design

Hierarchical generalization of language learners has typically been debated and studied within the *poverty of stimulus* (POS) paradigm (McCoy et al., 2020; Mueller et al., 2022; Ahuja et al., 2025; Mulligan et al., 2021; Kuribayashi and Baldwin, 2025): training data are curated intentionally ambiguous between a linear rule (**LINEAR**) and a hierarchical rule (**HIERARCHICAL**), and the model is then evaluated on held-out constructions where the two rules make divergent predictions, revealing which rule the model has adopted. When a number-mismatched noun—an *attractor*—intervenes between a plural subject and its verb (e.g., *Dogs on the donkey play/plays the cello*; Figure 2), **LINEAR** predicts agreement with the linearly nearest noun, whereas **HIERARCHICAL** predicts agreement with the syntactic subject. Attractor-bearing sentences thus serve as direct positive evidence of hierarchical rule (Linzen et al., 2016; Gulordava et al., 2018; Kuribayashi and Baldwin, 2025).

A growing body of work has identified factors that drive hierarchical generalization in LMs, including the amount of positive evidence in the training data (McCoy et al., 2020; Mueller and Linzen, 2023; Oba et al., 2024; Misra and Mahowald, 2024; Güven et al., 2025; Kuribayashi and Baldwin, 2025), the depth of clausal embedding (Qin et al., 2025), the choice of training objective (Ahuja et al., 2025), training duration beyond the point of overfitting (Murty et al., 2023), and access to additional modalities such as vision (Kuribayashi and Baldwin, 2025). By contrast, LMs trained on small, developmentally plausible inputs such as child-directed speech have been reported to fail to acquire the hierarchical rule (Yedetore et al., 2023; Padovani et al., 2025). All of these studies, however, share a monolingual design in which the language that provides positive evidence and the language on which generalization is evaluated coincide, leaving open whether the acquired hierarchical bias is tied to a particular language or instead reflects a language-general structural abstraction.

3 Tasks and Data

3.1 Cross-lingual Poverty-of-Stimulus (CPOS) Setting

The POS settings have typically been employed to analyze the inductive biases of LMs (Misra and Mahowald, 2024; Oba et al., 2024; Yang et al., 2026). The original POS argument in language acquisition in human infants claims the impossibility of identifying the correct linguistic rules underlying (infinite) linguistic input from finite observation, let alone limited exposure in human infants, necessitating strong inductive biases in human learners (Legate and Yang, 2002; Crain and Nakayama, 1987). Motivated by this argument, to assess the LMs' inductive biases, LMs are trained on data that are generally compatible with several generalization rules.[2] Controls on the availability of evidence during training and their evaluation clarify, for example, which rules LMs' inductive biases prefer and how sensitive LMs are to such evidence when inducing rules from data.

We extend this setting to be cross-lingual; positive evidence for a specific linguistic rule governing L1 and L2 grammars is only available in the L1 corpus. This clarifies the cross-lingual transfer ability of LMs by examining whether they can leverage L1 evidence for L2 generalization by controlling the number of instances with such evidence. Specifically, we first pretrain LMs on the L1 corpus with positive evidence, and continuously train them on

[2] Our study borrows the experimental settings of the POS paradigm and does not claim implications for the original POS debate.

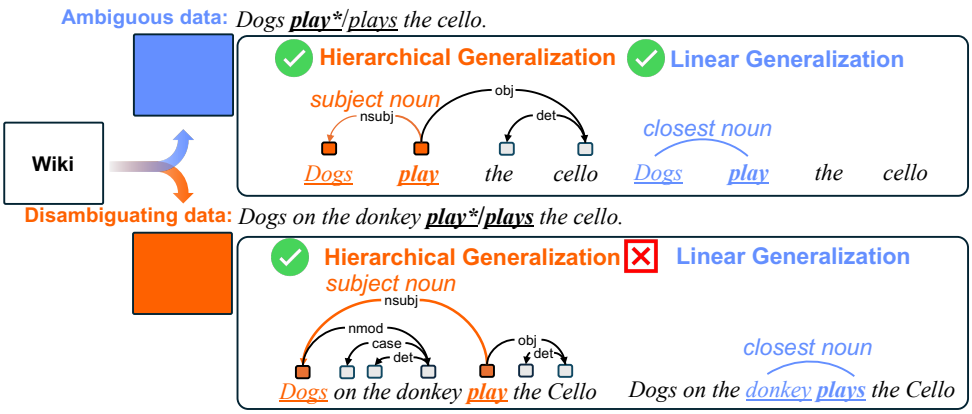

Figure 2: Example of **HIERARCHICAL** vs. **LINEAR** generalization for subject-verb number agreement. With an attractor (*donkey*) intervening between the plural subject (*Dogs*) and the verb, **LINEAR** predicts agreement with the linearly nearest noun (*plays*), whereas **HIERARCHICAL** predicts agreement with the syntactic subject (*play*).

the L2 corpus without such evidence, and evaluate the generalization on L2. We change the number of instances with certain positive evidence on the L1 side, and assess how sensitive the L2 generalization is to that L1 factor.

3.2 Evaluation Task (L2)

Following prior works (Linzen et al., 2016; Gulordava et al., 2018; Jumelet et al., 2026; Kuribayashi and Baldwin, 2025), we select the subject-verb agreement judgment as a testbed on the L2 side to evaluate the hierarchical generalization of LMs after cross-lingual transfer. The motivation behind this selection is that, first, as shown in Figure 1, one can clearly define instances (positive evidence) that disambiguate the hierarchical rule and the other ones that do not. Simply put, if the subject noun phrase (NP) is the closest NP to the verb, there are two possible generalizations: (i) the grammatical number of the verb should match that of its grammatical subject (**HIERARCHICAL** generalization), or (ii) it should match the closest NP to the verb, based on linear position (**LINEAR** generalization). The L1 training data should be controlled with respect to how informative the data are to disambiguate the **HIERARCHICAL** and **LINEAR** rules. For the evaluation, one can distinguish whether the model generalizes the agreement linearly or hierarchically, using examples with an attractor (e.g., *donkey* in Figure 1). Second, the subject-verb agreement phenomenon is widely attested in many languages, which enables scaling the experiments in multilingual settings, similarly to MultiBLiMP (Jumelet et al., 2026).

3.3 Positive Evidence for Hierarchical Generalization (L1)

We then define positive (direct and indirect) evidence for the hierarchical generalization of subject-verb agreement with two criteria. In the ablation experiments, the instances matching the following criteria are removed (or not) from the L1 corpus.

Attractors. Following Linzen et al. (2016) and Jumelet et al. (2026), if noun phrases (NPs) are embedded between the subject NP and its verb (attractors), and the grammatical number of the subject NP and attractor NPs differ, such an instance clarifies the agreement rule targeted in Section 3.2, serving as positive evidence. Note that there are several agreement phenomena similar to grammatical number, for example, grammatical gender agreement, which can also potentially disambiguate the hierarchical rules between NP and verb. It is difficult to enumerate and identify all features in different L1 language corpora, and we adopt a somewhat conservative criterion to remove all intervening NPs between NP and verb, regardless of their morphological features. Still, related constructions such as pre-/post-positional attachment can appear in many other constructions, and thus LMs are required to perform compositional generalization and identify hierarchical generalization rules.

Embedding depth. Another possible source of indirect evidence for the hierarchical generalization is not at the word–word level with morphological features, but also at the clause level; for example, center-embedding clarifies the hierarchical organization of sentence structure. To maximize the difference in the availability of cues for hierarchi-

cal generalization, we also remove such instances from the L1 corpus by counting the embedding depth of each sentence and filtering out the ones with an embedding depth of more than N.

3.4 L1 Training Data Creation

We create the L1 training data for four different, typologically diverse languages (i.e., English, German, Russian, and Japanese), and use Wikipedia[3] as a source of the corpora. We convert a base corpus into two versions of training data for each language, which differ only in the availability of positive evidence: a POS variant (**Ambiguous data**) and an evidence-enriched counterpart (**Disambiguating data**). Specifically, we first parse the sentences in the base corpus with Stanza (Qi et al., 2020) and classify each sentence into either a *positive-evidence sentence* that supports hierarchical generalization or a *non-evidential sentence* that supports both hierarchical and linear generalizations (Figure 2), according to the two criteria explained in Section 3.3, resulting in the following two versions of training data [4] with the same token counts:

Ambiguous data: the base corpus without all positive-evidence sentences, where a learner is exposed only to sentences that basically lack evidence for hierarchical generalization.

Disambiguating data: a corpus with positive evidence by replacing sentences in **Ambiguous data** with positive-evidence sentences.

We train different LMs by changing only their training data (Ambiguous data or Disambiguating data) and analyze the differences in their generalization patterns, which are attributed to the presence of positive evidence in L1. The implementation details of the parsing and classification are shown in the Appendix A.1.

4 Experimental Settings

We test the cross-lingual transfer of hierarchical generalization by manipulating two factors of L1 training: the level of exposure to **Disambiguating data**, and the choice of L1.

[3] https://huggingface.co/datasets/wikimedia/wikipedia

[4] Detailed statistics of the resulting training corpora are provided in Appendix A.2.

4.1 Condition 1: Exposure to Positive Evidence in L1

To assess whether L1 exposure to positive evidence drives transfer, we contrast two L1 training regimes that differ only in the availability of positive evidence during L1 training, adopting the following two scenarios:

HIER → LIN: L1 training data are sampled from **Disambiguating data**, and L2 training data from **Ambiguous data**.

LIN → LIN: both L1 and L2 training data are sampled from **Ambiguous data**; this serves as the baseline against which the **HIER → LIN** condition is compared.

4.2 Condition 2: Choice of L1

To assess whether transfer is modulated by typological properties of the source language, we vary L1 while holding L2 fixed at English. We select the following four languages that span a range of typological properties relevant to subject-verb agreement and the licensing of attractor configurations as L1:

L1 English Given that L2 is also English, this condition serves as a within-language reference for transfer.

L1 German German is a V2 language that, like English, licenses subject-verb agreement configurations involving attractors.

L1 Russian Russian also licenses attractor-bearing subject-verb agreement, but its considerably freer word order may weaken the attractor-as-evidence signal relative to English or German.

L1 Japanese Japanese exhibits very few attractor-bearing subject-verb agreement constructions.

These two conditions result in eight conditions in a {**HIER** → **LIN**, **LIN** → **LIN**} × {En, De, Ja, Ru} design.

4.3 Evaluation Metrics

MPP benchmarks. We evaluate LMs, following the minimal-pair paradigm (MPP), and use accuracy as our primary evaluation metric. For each minimal pair (s^+, s^-) in a benchmark $\mathcal{B}$, where s^+ is the grammatical sentence and s^- its minimally differing ungrammatical counterpart, we

count the item as correct when the model assigns lower sentence-level perplexity to s^+ than to s^-. Accuracy is the proportion of correct preferences over $\mathcal{B}$:

$$\text{acc} = \frac{1}{|\mathcal{B}|} \sum_{(s^+, s^-) \in \mathcal{B}} \mathbf{1}\big[\text{PPL}(s^+) < \text{PPL}(s^-)\big] \,. \tag{1}$$

We evaluate on two complementary benchmarks: MultiBLiMP (Jumelet et al., 2026), constructed from naturalistic multilingual corpora, and CLAMS (Mueller et al., 2020), a multilingual benchmark of synthetic minimal pairs. While we only use the English part of these benchmarks, we selected them as they target the subject-verb agreement phenomena. We denote the value of Equation (1) obtained under training condition c as acc_c. CLAMS provides a large, systematically controlled set of synthetic minimal pairs targeting attractor configurations, whereas MultiBLiMP comprises far fewer naturally occurring minimal pairs whose lexical frequency and collocational confounds inflate the variance of accuracy estimates. We therefore treat CLAMS as our primary benchmark and use MultiBLiMP as a directional check. We report, in addition to per-condition accuracy, the difference between the **HIER** →**LIN** and **LIN** →**LIN** conditions for each L1:

$$\Delta\text{acc} = \text{acc}_{\textbf{HIER}\rightarrow\textbf{LIN}} - \text{acc}_{\textbf{LIN}\rightarrow\textbf{LIN}} \;, \tag{2}$$

which isolates the contribution of L1 exposure to positive evidence; a positive Δacc is interpreted as evidence that the hierarchical generalization acquired in L1 has transferred to L2.

4.4 Detailed Setup

Training Data. Following the experimental design described in Section 3, we construct the L1 training data as a 50:50 mixture of **Disambiguating** and **Ambiguous** sentences, while sampling the L2 training data exclusively from the **Ambiguous** ones. In both L1 and L2 corpus, training and validation sets contain 50M and 5M tokens, respectively.

Models. All models adopt the GPT-2-small architecture (Radford et al., 2019), a standard Transformer language model trained with an autoregressive objective. Each model is trained from scratch on the L1 corpus for 8 epochs using this objective. After freezing[5] the transformer layers, pretraining continues on the L2 corpus for an additional 8 epochs, starting from the same checkpoint. Because each stage uses a 50M-token corpus, the model is exposed to a cumulative total of $50\text{M} \times 8 = 400\text{M}$ tokens per stage. Full hyperparameter settings are reported in Appendix C.

We train three models in each setting with different random seeds, and the results are averaged across these three runs.

Tokenizer. We train a single byte-level BPE tokenizer on the concatenation of the L1 and L2 corpora, with a vocabulary size of 64,000.

[5] We report results under the corresponding unfrozen setting in Appendix E.

L1	LIN → LIN	HIER → LIN
CLAMS		
En→En	56.04	61.49 (Δ +5.45)
De→En	52.83	54.52 (Δ +1.68)
Ru→En	51.07	52.11 (Δ +1.04)
Ja→En	51.24	51.77 (Δ +0.53)
MultiBLiMP		
En→En	61.39	69.92 (Δ +8.53)
De→En	60.45	57.91 (Δ -2.54)
Ru→En	60.45	61.02 (Δ +0.56)
Ja→En	58.47	60.45 (Δ +1.98)

Table 1: Accuracy (seed-mean of the final checkpoint, scored over tokens seen in L1 and/or L2 training data (OOV excluded)) under full vocabulary overlap, for benchmarks CLAMS, MultiBLiMP. The **HIER** → **LIN** column appends $\Delta = (\textbf{HIER} \rightarrow \textbf{LIN}) - (\textbf{LIN} \rightarrow \textbf{LIN})$ at the same overlap (green=positive, red=negative).

5 Experimental Results

We report the L2 (English) accuracy score in CLAMS and MultiBLiMP in Table 1; the corresponding learning trajectories through epochs are shown in Figure 3.

We first confirm whether L1 exposure to positive evidence yields above-chance L2 hierarchical-generalization accuracy. Under **LIN** → **LIN**, accuracy hovers around chance (51.1–56.0%) for all four L1 languages on CLAMS, confirming that under our design, the model fails to perform hierarchical generalization from **Ambiguous data** alone. By contrast, **HIER** → **LIN** yields $\Delta\text{acc} > 0$ for every L1, most prominently for English (+5.45), followed by German (+1.69) and Russian (+1.04; Table 1). As neither positive evidence is at the L2 learning phase, these gains must be attributed to the positive evidence provided in L1: linguistic evidence toward hierarchical generalization *does*, more or less, transfer across languages. The magni-

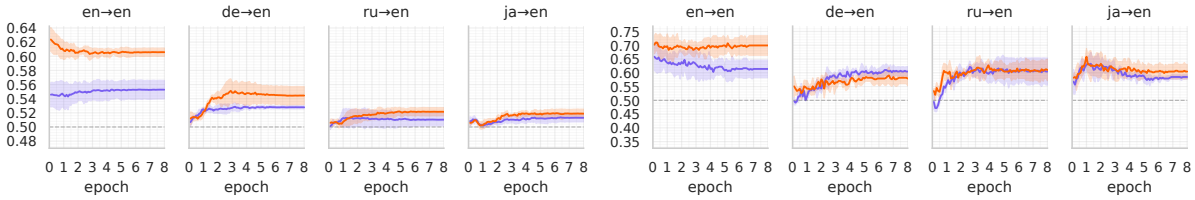

Figure 3: Learning trajectories on CLAMS (Left) and MultiBLiMP (Right) over L2-training epochs under the FULL OVERLAP condition. Orange lines: **HIER** → **LIN**; blue lines: **LIN** → **LIN** baseline. Lines show the seed mean; shaded bands indicate ±1 SD across three random seeds. The chance rate is 0.50.

tude of transfer varies with the L1 language. Δacc on CLAMS decreases monotonically from English (+5.45) > German (+1.68) > Russian (+1.04) > Japanese (+0.53). This gradient broadly aligns with the typological distance of each L1 from L2 English, which will be further clarified in §6.

On the MultiBLiMP dataset, English as L1 alone exhibits an even larger positive effect than on CLAMS (+8.53), reproducing the positive transfer evaluated on naturalistic data. The corresponding effects for German, Russian, and Japanese are −2.54, +0.57, and +1.98, respectively. MultiBLiMP contains far fewer items than CLAMS, which may bring somewhat conservative results. Notably, these transfer effects are not pronounced at any particular epoch in L2 training; rather, they become stable after several epochs (see Figure 3).

6 Analysis: Vocabulary Overlap

The previous experiments establish that LMs can, to a small extent, leverage the positive evidence observed in L1 data for L2 learning. However, these results alone do not establish that the observed differences arise specifically from the transfer of a hierarchical bias, as alternative explanations remain possible, including effects driven simply by lexical sharing between L1 and L2. We therefore extend our experiments by explicitly controlling vocabulary overlap and conduct statistical analysis.

Controlling Vocabulary Overlap We hypothesize that shared vocabulary likewise facilitates the transfer of hierarchical generalization by providing cross-lingual anchors through which an L1-acquired syntactic rule can be reused in L2; removing overlap should therefore suppress transfer.

To test this hypothesis, following Kallini et al. (2025), let V_{L_1} and V_{L_2} denote the sets of tokenizer tokens that occur in the L1 and L2 corpora, respectively, and let $O = V_{L_1} \cap V_{L_2}$ denote their native overlap. We control the subset $O' \subseteq O$ that retains shared token IDs across L1 and L2. See Table 6 for the resulting effective vocabulary sizes.

We compare three conditions: (i) the L1 and L2 tokenizer vocabularies are fully shared ($O' = O$), corresponding to the experiments up to §5 (FULL OVERLAP); (ii) only $|O'| = 1{,}000$ tokens of the native overlap are retained as shared, with the remainder offset, leaving only a minimal lexical bridge between the two languages (1K OVERLAP); and (iii) all tokens in O are offset ($O' = \emptyset$), so that L1 and L2 hold fully disjoint vocabulary sets (NO OVERLAP). That is, for example, in the En→En setting, the L2 English corpus is re-tokenized with a new tokenizer, and these IDs are appended to the L1 tokenizer as new IDs (e.g., *the* in L1 and L2 have different IDs). This is a somewhat extreme counterfactual baseline scenario; the learner learns a new language with perfectly the same syntactic rules as the L1, and only the vocabulary is different. The implementation details of disjoint and control overlap are shown in the Appendix B.

Statistical Analysis For each benchmark, we fit a Bayesian Bernoulli-logit mixed-effects model to item-level correctness across all 72 independently trained runs: `correct ~ L1_evidence * overlap * L1_language + (1|run) + (1|item)`; the full parameterization, priors, and convergence diagnostics are reported in Appendix D. The model estimates Δ_{hier} as the **HIER** →**LIN** minus **LIN** →**LIN** contrast within each overlap × L1-language combination; we report estimates (posterior medians) and equal-tailed 95% credible intervals (CrIs) on the log-odds scale, where positive estimates indicate higher odds of a correct L2 hierarchical-syntax preference after exposure to positive evidence in L1 and an effect is described as credible when its CrI excludes zero.

Results Table 2 reports cross-lingual transfer under each vocabulary overlap condition; Figure 4 shows the corresponding Δacc by L1 and epoch.

		LIN → LIN			HIER → LIN		
Data	L1→ L2	full	1K	no	full	1K	no
CLAMS	*En→En*	56.04	55.61	51.61	61.49 (Δ +5.45)	61.37 (Δ +5.76)	52.39 (Δ +0.79)
	De→En	52.83	52.78	51.42	54.52 (Δ +1.68)	55.46 (Δ +2.69)	52.19 (Δ +0.77)
	Ru→En	51.07	51.39	50.92	52.11 (Δ +1.04)	53.03 (Δ +1.64)	51.07 (Δ +0.14)
	Ja→En	51.24	51.62	51.85	51.77 (Δ +0.53)	51.75 (Δ +0.13)	51.59 (Δ -0.26)
MultiBLiMP	*En→En*	61.39	61.11	51.94	69.92 (Δ +8.53)	72.36 (Δ +11.25)	54.47 (Δ +2.53)
	De→En	60.45	59.04	57.34	57.91 (Δ -2.54)	57.91 (Δ -1.13)	56.21 (Δ -1.13)
	Ru→En	60.45	59.60	60.17	61.02 (Δ +0.56)	55.65 (Δ -3.95)	57.34 (Δ -2.82)
	Ja→En	58.47	59.89	60.73	60.45 (Δ +1.98)	59.04 (Δ -0.85)	60.73 (Δ +0.00)

Table 2: Accuracy (seed-mean of the final checkpoint, scored over tokens seen in L1 and/or L2 training data (OOV excluded)) under each vocabulary overlap conditions, for benchmarks CLAMS, MultiBLiMP. The **HIER → LIN** column appends $\Delta = (\textbf{HIER} \rightarrow \textbf{LIN}) - (\textbf{LIN} \rightarrow \textbf{LIN})$ at the same overlap (green=positive, red=negative).

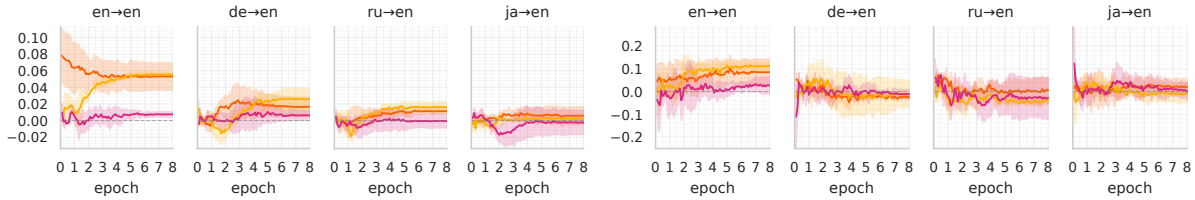

Figure 4: Δacc on CLAMS (Left) and MultiBLiMP (Right) broken down by L1, vocabulary-overlap condition, and L2-training epoch. Lines show the seed mean; shaded bands indicate ±1 SD across three random seeds.

Strikingly, whereas exposure to **Disambiguating data** in L1 yielded a reliable transfer under the FULL OVERLAP setting, the effect was almost entirely eliminated under the NO OVERLAP setting. Even in the En→En setting, the magnitude of Δ is drastically reduced. Under 1K OVERLAP, which retains only a minimal lexical bridge, the En→En case patterned with NO OVERLAP early in L2 training but recovered to a level on par with FULL OVERLAP by the late stages—a recovery that becomes especially clear when training is extended to 16 epochs (see Figures 5)[6]. On CLAMS, the mixed-effects analysis confirms this picture: the overall transfer contrast is credibly positive ($\Delta_{\text{hier}} = 0.121$ $[0.092, 0.150]$), and transfer is credibly stronger when vocabulary overlap is present than when it is absent ($\frac{\Delta_{\text{hier}}^{(\text{FULL})} + \Delta_{\text{hier}}^{(\text{1K})}}{2} - \Delta_{\text{hier}}^{(\text{NO})} = 0.144$ $[0.083, 0.205]$), whereas the contrast under NO OVERLAP is small and not credible ($\Delta_{\text{hier}}^{(\text{NO})} = 0.025$ $[-0.025, 0.074]$; Figure 6). At the L1 level on CLAMS, English and German settings show credible contrasts under vocabulary overlap, Japanese shows none, and under NO OVERLAP every credible interval includes zero (cell-level estimates in Table 9, Appendix D). On MultiBLiMP, the En→En contrast is credible under FULL OVERLAP and 1K OVERLAP but not under NO OVERLAP, while all non-English credible intervals include zero (Table 9); that is, the dependence of transfer on vocabulary overlap is not specific to the synthetic benchmark (CLAMS), but generally observed across several datasets.

[6]The results for other L1 languages are shown in Figure 7, 8

7 Discussion

Taken together, these observations indicate that cross-lingual transfer of hierarchical generalization *does* take place, but is tightly constrained by the amount of shared vocabulary, and that even a minimal overlap is sufficient for the model to exploit as an anchor for transfer.

8 Conclusion

In this paper, we extended the study of hierarchical generalization beyond its conventional monolingual setting by casting L1-to-L2 continued pretraining as a cross-lingual poverty-of-stimulus design in which positive evidence for the hierarchical rule is supplied only in L1. Across four typologically diverse L1s, models exposed to **Disambiguating data** during L1 training exhibited above-chance hierarchical accuracy at L2 even when the L2 input contained no positive evidence and the Transformer was frozen, with the magnitude of transfer varying systematically across L1 languages. Ma-

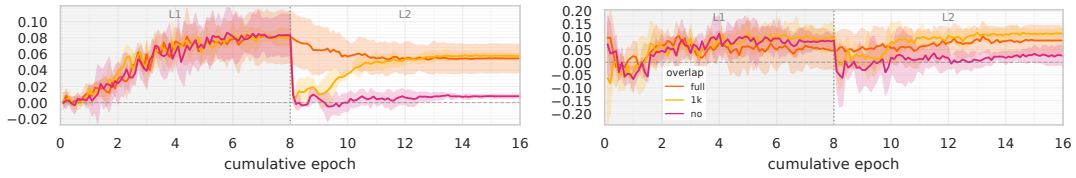

Figure 5: Δacc on CLAMS(Left) and MultiBLiMP(Right) for En→En under extended training (8epoch in L1→ 8epoch in L2). Lines show the seed mean; shaded bands indicate±1 SD across three random seeds.

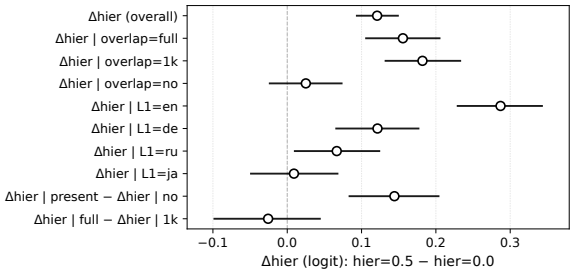

Figure 6: Principal transfer contrasts on CLAMS from the mixed-effects model. Points are posterior medians of Δ_{hier} (log-odds) and error bars are 95% CrIs; the dashed line marks zero. In the axis labels, hier=0.5 and hier=0.0 correspond to the **HIER** →**LIN** and **LIN** →**LIN** conditions, respectively.

nipulating the L1–L2 shared-vocabulary inventory revealed that this transfer is tightly mediated by lexical overlap: making the two vocabularies fully disjoint nearly eliminated the effect, whereas a bridge of as few as 1,000 shared tokens sufficed for the late-training effect to recover to fully shared levels. Taken together, these findings indicate that the inductive bias for hierarchical generalization acquired during pretraining is not strictly tied to a single language, yet its deployment in a new language is conditional on a lexical bridge between the two—a finding that bears directly on the design of multilingual tokenizers when cross-lingual reuse of syntactic abstractions is at stake.

Limitations

First, our experimental scope is limited in model size and language sample. All experiments use GPT-2-small; larger models and alternative architectures are not tested. Following Arnett et al. (2025), we take the position that a phenomenon is best established at small scale before its scaling behaviour is examined. We further restrict L1 to four languages and fix L2 to English, chosen to span a gradient of word-order flexibility and attractor-licensing frequency; whether the findings generalize to typologically more diverse pairs, particularly outside Indo-European, remains untested. In particular, replicating the language gradient with L2 varied would be necessary to disentangle L1/L2 status from pair-specific lexical and structural proximity.

Second, the design controls that strengthen interpretability also constrain external validity. Freezing the Transformer during L2 training is what lets us isolate transfer of L1-acquired generalization from embedding-layer adaptation; our findings therefore do not extrapolate directly to standard multilingual pretraining or joint-training regimes, in which the Transformer is updated simultaneously. Whether the dependence on vocabulary overlap survives more flexible fine-tuning remains open. We also operationalize hierarchy through attractor-based subject-verb number agreement and clausal embedding depth, following the standard operationalization since Linzen et al. (2016); transfer on other hierarchical phenomena such as filler–gap dependencies or island constraints lies outside our present scope.

Third, synthetic and naturalistic evaluations diverge. The language gradient in §5 is observed on CLAMS, but on MultiBLiMP a credible effect appears only for English, with the sign of the German point estimate reversed. At least three accounts are compatible with this pattern: (i) the gradient is an artefact of statistical regularities specific to synthetic data; (ii) residual biases in naturalistic data—lexical frequency, collocational strength—mask the contribution of hierarchical generalization; or (iii) the smaller item counts of MultiBLiMP, together with three seeds, raise the minimal detectable effect beyond what our power supports. The fact that the English effect is reproduced on MultiBLiMP is compatible with (ii)+(iii), but our data do not adjudicate among the three. Disentangling them will require naturalistic benchmarks with larger item counts and replications with more seeds.

Ethical Statement

We use AI services, including ChatGPT, Gemini, and Claude, for proofreading and writing assistance. Furthermore, we manually review the generated text to confirm that it contains no hallucinations, false statements, or reward-hacking code. Ultimately, the authors remain fully accountable for the accuracy and integrity of the entire text. It is a study of language and language acquisition, and we did not intentionally generate or promote harmful content in our experiments.

Acknowledgements

We would like to express our gratitude for the anonymous reviewers who provided many insightful comments that have improved our paper. This work was supported by JSPS KAKENHI Grant Numbers 23K16938, 24K16077 and 26H02511.

References

Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A. Smith, Navin Goyal, and Yulia Tsvetkov. 2025. Learning syntax without planting trees: Understanding hierarchical generalization in transformers. *Transactions of the Association for Computational Linguistics*, 13:121–141.

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.

Catherine Arnett, Tyler A. Chang, James A. Michaelov, and Ben Bergen. 2025. On the acquisition of shared grammatical representations in bilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20707–20726, Vienna, Austria. Association for Computational Linguistics.

Damián E. Blasi, Ryan Cotterell, Lawrence Wolf-Sonkin, Sabine Stoll, Balthasar Bickel, and Marco Baroni. 2019. On the distribution of deep clausal embeddings: A large cross-linguistic study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3938–3943, Florence, Italy. Association for Computational Linguistics.

Noam Chomsky. 1980. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating Critical Period Effects in Language Acquisition through Neural Language Models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Stephen Crain and Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language*, 63(3):522–543.

Yuan Gao, Suchir Salhan, Andrew Caines, Paula Buttery, and Weiwei Sun. 2025. BLiSS: Evaluating bilingual learner competence in second language small language models. In *Proceedings of the First BabyLM Workshop*, pages 160–174, Suzhou, China. Association for Computational Linguistics.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. The MIT Press.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Arzu Burcu Güven, Anna Rogers, and Rob Van Der Goot. 2025. Do syntactic categories help in developmentally motivated curriculum learning for language models? In *Proceedings of the First BabyLM Workshop*, pages 288–300, Suzhou, China. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False Friends are not foes: Investigating vocabulary overlap in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21138–21154, Suzhou, China. Association for Computational Linguistics.

Tatsuki Kuribayashi and Timothy Baldwin. 2025. Does vision accelerate hierarchical generalization in neural language learners? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1865–1879, Abu Dhabi, UAE. Association for Computational Linguistics.

Julie Anne Legate and Charles D Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1-2):151–162.

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.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland. Association for Computational Linguistics.

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252, Toronto, Canada. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Karl Mulligan, Robert Frank, and Tal Linzen. 2021. Structure here, bias there: Hierarchical generalization by jointly learning syntactic transformations. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 125–135, Online. Association for Computational Linguistics.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. Grokking of hierarchical structure in vanilla transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada. Association for Computational Linguistics.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Miyu Oba, Yohei Oseki, Akiyo Fukatsu, Akari Haga, Hiroki Ouchi, Taro Watanabe, and Saku Sugawara. 2024. Can language models induce grammatical knowledge from indirect evidence? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20591–20603, Miami, Florida, USA. Association for Computational Linguistics.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025. Child-directed language does not consistently boost syntax learning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19735–19756, Suzhou, China. 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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Tian Qin, Naomi Saphra, and David Alvarez-Melis. 2025. Data drives unstable hierarchical generalization in LMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11722–11740, Suzhou, China. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

Aditya Yadavalli, Alekhya Yadavalli, and 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, Toronto, Canada. Association for Computational Linguistics.

Xiulin Yang, Arianna Bisazza, Nathan Schneider, and Ethan Gotlieb Wilcox. 2026. A unified assessment of the poverty of the stimulus argument for neural language models. *Preprint*, arXiv:2602.09992.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

Iza Škrjanec, Irene Elisabeth Winther, Vera Demberg, and Stefan L. Frank. 2026. Is cross-lingual transfer in bilingual models human-like? a study with overlapping word forms in dutch and english. *Preprint*, arXiv:2604.07067.

A Supplementary Material

A.1 Corpus Construction Details

We use Stanza (Qi et al., 2020) both for token counting and positive evidence identification[7]. The training data are drawn from Wikipedia/Wikimedia because (1) its domain closely matches that of the Universal Dependencies treebanks on which Stanza's parser was trained, yielding more reliable parses than developmentally plausible alternatives such as the child-directed speech in CHILDES, and (2) it is available at comparable scale across our four target languages, a prerequisite for the matched-size manipulation described in Section 3.4.

After parsing, we classify every sentence containing an `nsubj` relation along two features that follow the criteria in Section 3.3. `has_attractors` (bool), coming from the implementation of Jumelet et al. (2026), indicates whether any noun appears linearly between the subject and the verb of the `nsubj` relation. Following the clausal embedding in Blasi et al. (2019), `embedding_depth` (int) is computed by traversing the head path from each word to the sentence root and incrementing the count whenever the path crosses a dependency relation that marks an embedded clause (`acl`, `advcl`, `ccomp`, `xcomp`, `csubj`, and their subtypes); the sentence-level `embedding_depth` is the maximum of these per-word depths.

Classification criteria. Sentences can be classified into one of the following two types.

non-evidential sentence: Sentences that do not contain attractors and have an embedding depth of 0.

positive-evidence sentence: Sentences that contain attractors and have an embedding depth of at least 1.

A.2 Training Corpus Statistics

Table 3 reports the per-language sentence counts and average sentence lengths for the **Ambiguous data** and **Disambiguating data** partitions of the training corpora.

[7]The performance of the pretrained Stanza models for all four languages is publicly reported on the official model-performance page. https://stanfordnlp.github.io/stanza/performance.html

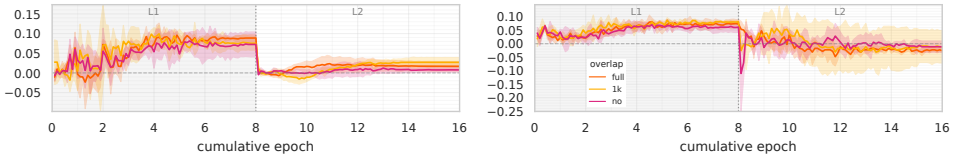

Figure 7: Δacc on CLAMS(Left) and MultiBLiMP(Right) for De→En under extended training (8epoch in L1→ 8epoch in L2). Lines show the seed mean; shaded bands indicate±1 SD across three random seeds.

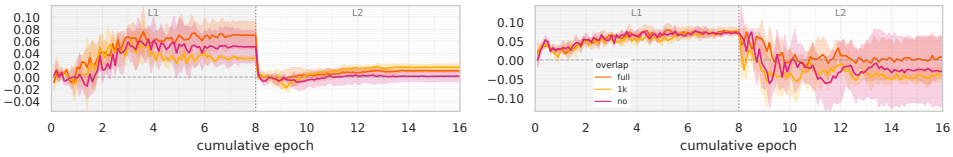

Figure 8: Δacc on CLAMS(Left) and MultiBLiMP(Right) for Ru→En under extended training (8epoch in L1→ 8epoch in L2). Lines show the seed mean; shaded bands indicate±1 SD across three random seeds.

A.3 Evaluation Benchmark Coverage

Table 5 lists the phenomena included in the hierarchical partitions of CLAMS and MultiBLiMP, together with the number of minimal-pair items per phenomenon.

B Details of Vocabulary Overlap

B.1 Implementation of Vocabulary Overlap.

Following the implementation of Kallini et al. (2025), we take the byte-level BPE tokenizer $\mathcal{T}$ trained on the concatenation of L1 and L2 (Section 4.4) as our base and extract the native overlap $O = V_{L_1} \cap V_{L_2}$, i.e. the set of tokens that appear in both the L1 and L2 corpora. Each overlap condition is fully specified by the subset $O' \subseteq O$ that is retained as shared: tokens in $O \setminus O'$ are remapped on the L2 side by an offset of $|O|$ so as to occupy a disjoint ID range, while L2-specific (non-shared) tokens keep their original IDs. Crucially, the offset amount is the shared-vocabulary size $|O|$ rather than the base vocabulary size N, ensuring that only the shared portion is displaced into the disjoint region 9. Thus, FULL OVERLAP corresponds to $O' = O$ (no offset), NO OVERLAP to $O' = \emptyset$ (all of O offset), and 1K OVERLAP to $|O'| = 1{,}000$, where $|O| - 1{,}000$ shared tokens are offset and 1,000 remain shared as a minimal lexical anchor between the two languages. The resulting L1- and L2-specific effective vocabulary sizes under each condition are summarized in Table 6. Model architecture, hyperparameters, and the training/evaluation procedure are identical to those in §4.4.

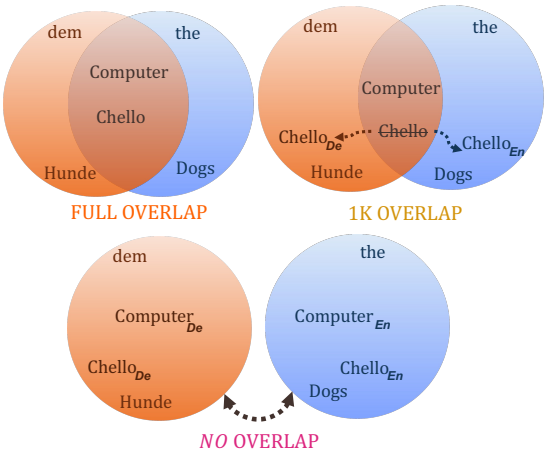

Figure 9: A visualization of the three overlap conditions. FULL OVERLAP, NO OVERLAP, 1K OVERLAP.

B.2 Vocabulary Overlap and Parameter Counts

Table 6 reports the L1- and L2-specific effective vocabulary sizes and total parameter counts for each (language pair, overlap setting) combination, separately for the L1 **Ambiguous data** and **Disambiguating data**training corpora.

B.3 Coverage of Overlap vocabulary in the Training Corpus

Table 7 reports the coverage of shared and disjoint vocabulary under each condition of vocabulary overlap.

C Hyperparameters

Table 8 reports the architectural and optimization hyperparameters used in all experiments. The same configuration is used across L1 and L2 training, the

Language	Ambiguous data		Disambiguating data	
	#Sents	#Words/Sent	#Sents	#Words/Sent
English (L1/L2)	2.42/2.42M	20.6/20.6	2.0M	24.6
German	2.7M	18.3	2.3M	21.9
Russian	2.8M	17.8	2.3M	21.3
Japanese	3.8M	13.1	2.8M	17.9

Table 3: Statistics of the training corpora. Each language group reports the number of sentences and the average sentence length for the **Ambiguous data** and **Disambiguating data** partitions. The English row additionally reports the L2 column since English serves as the L2 in all experiments.

Language	Ambiguous data					Disambiguating data				
	nsubj	Clausal	All	Depth	DLT Stor.	nsubj	Clausal	All	Depth	DLT Stor.
English	1.64	0.00	3.29	0.00	4.49	4.33	5.70	3.13	0.62	4.93
German	1.38	0.00	3.21	0.00	4.24	5.44	10.64	3.47	0.48	5.38
Russian	1.46	0.00	3.31	0.00	3.90	4.14	5.37	2.97	0.53	4.36
Japanese	2.30	0.00	2.87	0.00	2.55	7.06	5.27	2.88	0.66	2.96

Table 4: Dependency-based complexity statistics of the 50M-token L1 training corpora: nsubj = mean subject–verb linear distance over `nsubj` relations; Clausal = mean head–dependent distance over clausal relations (`acl`, `advcl`, `ccomp`, `xcomp`, `csubj`); All = the same over all relations except `punct`; Depth = mean per-sentence maximum clausal embedding depth; DLT Stor. = mean per-sentence maximum storage cost under the Dependency Locality Theory (Gibson, 2000), i.e., the number of pending syntactic heads. Clausal and Depth are exactly 0 in the **LIN** →**LIN** corpora by construction. The English L2 corpus matches the English **LIN** →**LIN** columns to two decimals.

Phenomenon (subcategory)	**# items**
CLAMS (4 phenomena)	
`obj_rel_across_anim`	11,200
`obj_rel_within_anim`	11,200
`prep_anim`	16,800
`subj_rel`	11,200
Total	*50,400*
MultiBLiMP (1 phenomenon)	
`has_attractors_true`	130
Total	*130*

Table 5: Phenomena included in the *hierarchical* partition of CLAMS and MultiBLiMP, together with the number of minimal-pair items per phenomenon. These are the items aggregated into the accuracy reported in Section 5.

only difference being that L2 training freezes the Transformer layers, as described in Section 4.4.

D Details of the Statistical Analysis

We fit a separate Bayesian Bernoulli-logit mixed-effects model for each benchmark using all 72 independently trained runs (2 L1-evidence conditions × 3 overlap conditions × 4 L1 languages × 3 seeds). For response y_n in cell $c[n]$, item $i[n]$, and run $r[n]$, the model is $y_n \sim \text{Bernoulli}(\text{logit}^{-1}(\mu_{c[n]} + a_{i[n]} + b_{r[n]}))$, where c indexes the L1-evidence × overlap × L1-language cells. We use cell-means parameterization of the fully interacted model in Section 6, with one mean for each of the 24 L1-evidence × overlap × L1-language combinations; Δ_{hier} is the **HIER** →**LIN** minus **LIN** →**LIN** contrast between the corresponding posterior cell means. The item and run intercepts account for repeated evaluation of the same item across conditions and dependence among items evaluated with the same trained model, respectively, where a run is one L1-evidence × overlap × L1-language × seed realization. We use weakly informative priors, with $\mu_c \sim \text{Normal}(0, 2.5)$, $a_i \sim \text{Normal}(0, \sigma_{\text{item}})$, $b_r \sim \text{Normal}(0, \sigma_{\text{run}})$, $\sigma_{\text{item}} \sim \text{HalfNormal}(1.0)$, and $\sigma_{\text{run}} \sim \text{HalfNormal}(0.5)$. Sampling used four chains with 4,000 posterior draws each; there were no divergent transitions and all $\hat{R} \leq 1.01$.

E Unfreezing Experiment

To test whether freezing the Transformer during L2 training drives the vocabulary-overlap effect, we keep the remaining training and evaluation settings identical to the main experiment and change only whether the Transformer layers are updated during L2 training (Figure 10). The ablation compares the FULL OVERLAP and NO OVERLAP conditions.

Language Pair	Setting	$\lvert V_1 \rvert$ Lin	$\lvert V_1 \rvert$ Hier	$\lvert V_2 \rvert$ Lin	$\lvert V_2 \rvert$ Hier	$\lvert O' \rvert$ Lin	$\lvert O' \rvert$ Hier	**Vocabulary Size** Lin	**Vocabulary Size** Hier	**Total Parameters (M)** Lin	**Total Parameters (M)** Hier
En→En	Full Overlap	63,839	63,789	63,825	63,820	63,768	63,733	64,000	64,000	134.60	134.60
	1k Overlap					1,000	1,000	126,768	126,733	182.81	182.78
	No Overlap					0	0	127,768	127,733	183.58	183.55
De→En	Full Overlap	62,676	62,238	56,707	55,048	55,473	53,374	64,000	64,000	134.60	134.60
	1k Overlap					1,000	1,000	118,473	116,374	176.44	174.82
	No Overlap					0	0	119,473	117,374	177.20	175.59
Ru→En	Full Overlap	62,463	61,936	38,177	36,306	36,742	34,334	64,000	64,000	134.60	134.60
	1k Overlap					1,000	1,000	99,742	97,334	162.05	160.20
	No Overlap					0	0	100,742	98,334	162.82	160.97
Ja→En	Full Overlap	63,015	62,336	37,487	37,862	36,665	36,354	64,000	64,000	134.60	134.60
	1k Overlap					1,000	1,000	99,665	99,354	161.99	161.75
	No Overlap					0	0	100,665	100,354	162.76	162.52

Table 6: Token statistics, vocabulary sizes, and parameter counts for each (language pair, overlap setting) combination, shown separately for the L1 **LIN** (*hier_ratio*=0.0) and **HIER** (*hier_ratio*=0.5) training corpora. $|V_1|$ and $|V_2|$ denote the L1- and L2-corpus vocabularies; $|O'|$ is the number of shared tokens retained after the disjoint strategy. Total parameters (in millions) are computed analytically from $\text{vocab_size} \times n_{\text{embd}} + n_{\text{pos}} \times n_{\text{embd}} + n_{\text{layer}} \times (12\, n_{\text{embd}}^2 + 13\, n_{\text{embd}}) + 2\, n_{\text{embd}}$.

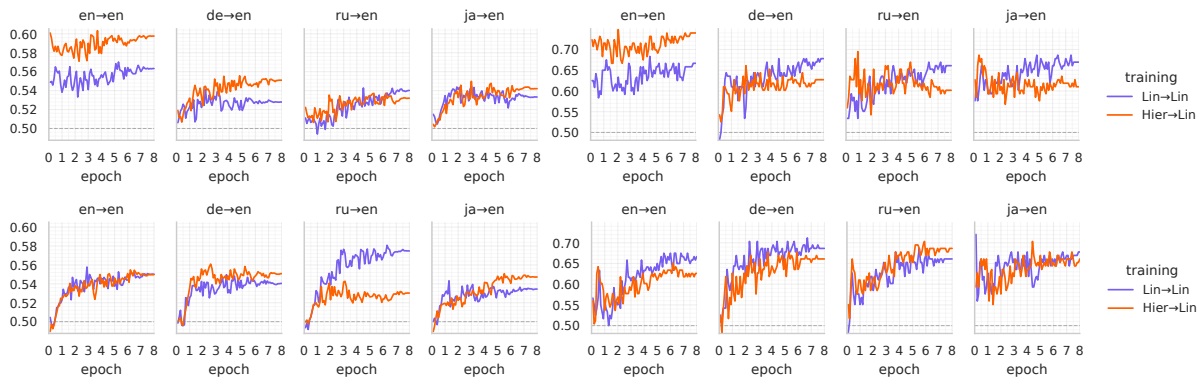

Figure 10: Results of the unfrozen-Transformer ablation under FULL OVERLAP (top) and NO OVERLAP (bottom); all Transformer layers are updated during L2 training, while the remaining settings match the main experiment.

F Computational Resources

Across all experiments, we used a total of approximately 1116 GPU hours. Training was conducted on NVIDIA A100 80GB and NVIDIA A100 40GB GPUs.

Language Pair	Setting	L1 corpus tokens (%)						L2 corpus tokens (%)					
		Shared		Disjointed		L1-unique		Shared		Disjointed		L2-unique	
		Lin	Hier	Lin	Hier	Lin	Hier	Lin	Hier	Lin	Hier	Lin	Hier
En→En	Full Overlap	100.00	100.00	0.00	0.00	0.00	0.00	100.00	99.99	0.00	0.00	0.00	0.01
	1k Overlap	60.06	59.26	39.94	40.74			60.05	59.98	39.94	40.01		
	No Overlap	0.00	0.00	100.00	100.00			0.00	0.00	100.00	99.99		
De→En	Full Overlap	94.32	92.56	0.00	0.00	5.68	7.44	99.04	98.70	0.00	0.00	0.96	1.30
	1k Overlap	43.44	39.32	50.87	53.24			50.94	50.96	48.10	47.74		
	No Overlap	0.00	0.00	94.32	92.56			0.00	0.00	99.04	98.70		
Ru→En	Full Overlap	71.12	66.70	0.00	0.00	28.88	33.30	98.75	98.22	0.00	0.00	1.25	1.78
	1k Overlap	27.80	22.60	43.32	44.10			49.74	48.96	49.01	49.26		
	No Overlap	0.00	0.00	71.12	66.70			0.00	0.00	98.75	98.22		
Ja→En	Full Overlap	76.42	73.27	0.00	0.00	23.58	26.73	99.18	98.62	0.00	0.00	0.82	1.38
	1k Overlap	17.74	12.93	58.68	60.34			49.71	48.59	49.47	50.03		
	No Overlap	0.00	0.00	76.42	73.27			0.00	0.00	99.18	98.62		

Table 7: Per-corpus token coverage (%) of the four dissociation categories. Cells are the share of the 50M-word L1 (resp. L2) training corpus whose tokens fall into each category, computed by tokenizing each corpus with its (language pair, hier_ratio)-specific joint Byte-Level BPE. Categories: *Shared* are intersection tokens whose vocabulary IDs are kept identical across L1 and L2; *Disjointed* are intersection tokens whose L2-side IDs are shifted into the $[V, V+D)$ range, giving them separate embeddings; *L1- / L2-unique* are tokens that appear only in the respective corpus. L1-unique on the L1 side (and L2-unique on the L2 side) are constant across overlap settings within a (pair, hier) cell and shown via `multirow`; the complementary cells (L2-unique on the L1 side and vice versa) are zero by definition and omitted. Within each (pair, hier) the three pct values (Shared, Disjointed, language-specific unique) sum to $\approx 100\%$ on each side.

Hyperparameter	Setting
Training-corpus size per stage	50M tokens
Number of layers (n_{layer})	12
Number of attention heads (n_{head})	12
Hidden size (n_{embd})	768
learning rate	5e-4
Per-device batch size	32
Gradient accumulation steps	2
context length	512
warmup steps	1,500
Epochs per stage	8
dropout	0.1
weight decay	0.1
learning rate scheduler	cosine
Evaluation	
Checkpoint save interval	every 1 epoch
Validation interval	every 0.05 epoch
Benchmark evaluation interval	every 0.1 epoch

Table 8: Hyperparameter settings used for the experiments.

L1	Overlap	Estimate	95% CrI
	FULL	**0.390**	[**0.289**, **0.489**]
En	1K	**0.414**	[**0.311**, **0.520**]
	NO	0.055	[−0.042, 0.156]
	FULL	**0.119**	[**0.017**, **0.221**]
De	1K	**0.191**	[**0.089**, **0.293**]
	NO	0.054	[−0.045, 0.152]
	FULL	0.076	[−0.024, 0.174]
Ru	1K	**0.115**	[**0.013**, **0.216**]
	NO	0.010	[−0.089, 0.108]
	FULL	0.039	[−0.065, 0.141]
Ja	1K	0.007	[−0.092, 0.110]
	NO	−0.018	[−0.122, 0.085]
Overlap-presence interaction by L1			
En	Present − No	**0.346**	[**0.220**, **0.470**]
De	Present − No	0.102	[−0.021, 0.222]
Ru	Present − No	0.085	[−0.036, 0.208]
Ja	Present − No	0.042	[−0.083, 0.168]

(a) CLAMS

L1	Overlap	Estimate	95% CrI
	FULL	**0.627**	[**0.217**, **1.049**]
En	1K	**0.840**	[**0.425**, **1.259**]
	NO	0.188	[−0.217, 0.596]
	FULL	−0.160	[−0.562, 0.249]
De	1K	−0.081	[−0.486, 0.328]
	NO	−0.078	[−0.481, 0.334]
	FULL	0.018	[−0.388, 0.429]
Ru	1K	−0.279	[−0.690, 0.131]
	NO	−0.199	[−0.602, 0.210]
	FULL	0.141	[−0.268, 0.538]
Ja	1K	−0.060	[−0.468, 0.348]
	NO	0.039	[−0.366, 0.444]
Overlap-presence interaction by L1			
En	Present − No	**0.545**	[**0.047**, **1.048**]
De	Present − No	−0.045	[−0.539, 0.449]
Ru	Present − No	0.070	[−0.425, 0.562]
Ja	Present − No	0.003	[−0.505, 0.505]

(b) MultiBLiMP

Table 9: Model-estimated transfer contrasts and planned overlap-presence difference-in-differences by L1. Rows report the model estimate of Δ_{hier} within each vocabulary-overlap $\times$ L1 combination; "Present − No" denotes $(\Delta_{\text{FULL}} + \Delta_{\text{1K}})/2 - \Delta_{\text{NO}}$. Entries are estimates (posterior medians) with equal-tailed 95% CrIs on the log-odds scale, and positive estimates indicate a higher probability of a correct response under **HIER →LIN** than under **LIN →LIN**. **Bold** entries have CrIs that lie entirely above zero and are therefore described as credibly positive.