On the Acquisition of Shared Grammatical Representations in Bilingual Language Models

Catherine Arnett[a,b]   **Tyler A. Chang**[c]   **James A. Michaelov**[c,d,e]   **Benjamin K. Bergen**[c]

[a]Department of Linguistics, UCSD   [b]EleutherAI

[c]Department of Cognitive Science, UCSD

[d]Department of Brain and Cognitive Sciences, MIT   [e]MIT Libraries CREOS

catherine@eleuther.ai, tachang@ucsd.edu, jamic@mit.edu, bkbergen@ucsd.edu

Abstract

Crosslingual transfer is crucial to contemporary language models' multilingual capabilities, but how it occurs is not well understood. We ask what happens to a monolingual language model when it begins to be trained on a second language. Specifically, we train small bilingual models for which we control the amount of data for each language and the order of language exposure. To find evidence of shared multilingual representations, we turn to structural priming, a method used to study grammatical representations in humans. We first replicate previous crosslingual structural priming results and find that after controlling for training data quantity and language exposure, there are asymmetrical effects across language pairs and directions. We argue that this asymmetry may shape hypotheses about human structural priming effects. We also find that structural priming effects are less robust for less similar language pairs, highlighting potential limitations of crosslingual transfer learning and shared representations for typologically diverse languages.

🤗 B-GPT models   ✳ code and data

1 Introduction

Multilingual language models share representations across languages (Artetxe et al., 2020; Conneau et al., 2020), which is thought to be crucial for crosslingual transfer abilities (Wu and Dredze, 2019; Chi et al., 2020; Hu et al., 2020; Winata et al., 2021, 2022). While there has been much evidence that successful crosslingual transfer can enable improvements in performance, there has not yet been extensive research about how models develop the shared representations that drive it.

To study these shared representations, we use structural priming, a phenomenon in which a target sentence with a congruent preceding prime sentence type will have a higher likelihood than the same target sentence following an incongruent prime. For example, we predict a language

model would assign a higher probability to a prepositional object (PO) dative sentence (e.g. "*the chef gives a hat to the swimmer*") following another PO sentence than following a double object (DO) dative sentence (e.g. "*the chef gives the swimmer a hat*"; sentences from Schoonbaert et al., 2007). In crosslingual structural priming, targets that share a grammatical construction with the prime are more likely, even if the two sentences are in different languages (Figure 1).

Human experiments demonstrate robust structural priming effects in a wide variety of languages (Bock, 1986; see Pickering and Ferreira (2008) for review) and have been used to argue that bilinguals have shared grammatical representations for their languages. Structural priming has previously been used to study the structural representations learned by language models (Prasad et al., 2019; Sinclair et al., 2022; Frank, 2021; Li et al., 2022; Choi and Park, 2022; Michaelov et al., 2023; Jumelet et al., 2024; Zhou et al., 2024). Because the grammatical structure is primed rather than a specific semantic meaning, Sinclair et al. (2022) argue that structural priming effects provide evidence for abstract grammatical representations in language models. By measuring output model probabilities given a prime sentence, structural priming demonstrates causal effects of grammatical representations on model outputs without relying on access to internal model states. The presence of structural priming in crosslingual scenarios (e.g. a structure primes a similar structure in another language) would indicate representations shared between languages.

Michaelov et al. (2023) provided the first evidence for crosslingual structural priming in Transformer language models. The authors argued that this was evidence that language models use shared abstract grammatical representations to represent grammatical constructions for multiple languages. However, they reported variable and asymmetric effects where for some pairs of languages, struc-
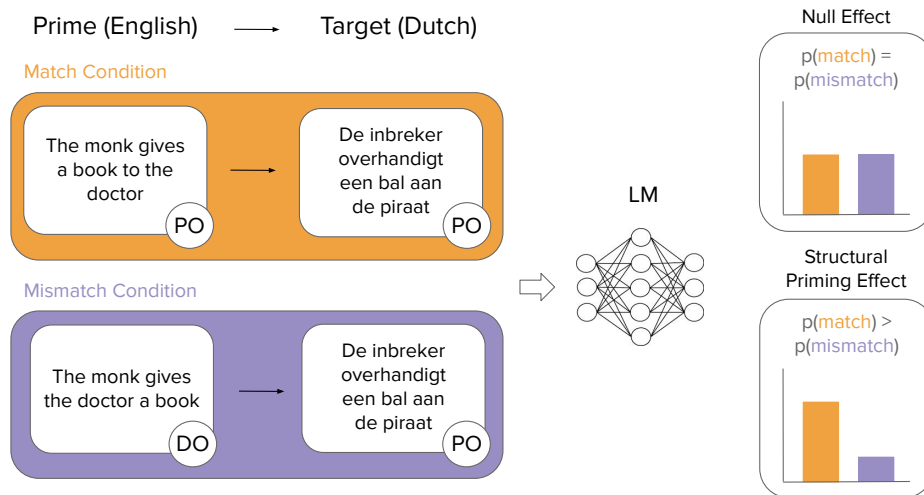
Figure 1: Structural priming paradigm.

tural priming effects were stronger in one direction and weaker (or even non-existent) in the other.

Language Asymmetries In this paper, we first investigate why there are asymmetric effects between languages, i.e. depending on whether they are the target or prime language. Michaelov et al. (2023) observe that structural priming effects are stronger when the target language is English. The same effect, observed in human experiments, has been attributed to which language is the first or second learned language (L1 or L2). It is generally thought that structural priming effects are stronger when the prime language is L1 and the target language is L2 (henceforth L1→L2 priming; Schoonbaert et al., 2007). However, a major confound in this line of research is that in most psycholinguistic experiments, English is the L2. This is due to the populations which are usually sampled from for these experiments, such as university students in countries like the Netherlands (e.g. Schoonbaert et al., 2007; Bernolet et al., 2013), where it is easiest to find L1 Dutch and L2 English speakers. If we train language models with different orders of language exposure do we see the same asymmetries? If we do, then the asymmetries are not necessarily due to L1 versus L2, but instead the target language itself (which in this case is English).

Language Similarity Second, we investigate whether language similarity impacts structural priming effects. In Michaelov et al. (2023), we found more robust structural priming effects for English-Dutch and English-Spanish than for English-Polish and English-Greek sentence pairs.

We speculated that this could be in part due to the lower proportions of Polish and Greek training data in the models tested. However, it could also be due to varying language similarity; crosslingual transfer has been shown to be more effective between more similar languages (Lin et al., 2019; Ogueji et al., 2021; Chang et al., 2024a), suggesting a greater degree of representation sharing in similar languages. Polish and Greek are typologically less similar to English than Dutch and Spanish are (§5.2), which might lead to weaker crosslingual structural priming effects. After controlling for the amount of pre-training data, do differences in the robustness of structural priming correspond to differences in language similarity? If so, then it is possible that asymmetries are not due to order or language exposure, but instead about features of either the prime or target language.

Training Dynamics Previous structural priming studies involving language models focus on the final model checkpoint. Here, we ask whether there is a temporal link between the model's acquisition of grammatical knowledge in a second language and its exhibiting structural priming effects. If structural priming effects emerge only after the model learns non-trivial grammatical representations, this reinforces the value of structural priming as a tool for studying multilingual representations.

Contributions We train bilingual models, varying the amount of data for each language and the order in which the language model is exposed to each language. With these models, we replicate previous structural priming experiments. We find that asym-

metries persist, even when the order of language exposure is reversed. This suggests that asymmetries may be due to features of the target language. Our models also show more robust structural priming effects for more similar language pairs. Together, our results not only shed light on shared representations in language models, but may inform our understanding of human structural priming effects.

2 Related Work

Language Models as Model Organisms Our work relates to an ongoing discussion about the role of language models in linguistics and cognitive science (Piantadosi, 2023; Mahowald et al., 2024; Futrell and Mahowald, 2025). In a sense, language models are the first *model organism* for language researchers (akin to fruit flies in genetics research), in that they offer the possibility to refine hypotheses about language through the manipulation and evaluation of models, with direct or indirect implications for linguistic theory and related disciplines (Müller, 2024). For example, in neurolinguistics, Jain et al. (2024) argue that such *in silico* testing is valuable for evaluating construct validity and refining experiments before they are conducted, as neurolinguistic experiments are extremely costly to run. Similarly, recent work has shown that language models can be valuable model organisms for questions where controlled manipulations are not possible in human experiments. Recent work has used manipulations of training data, for example removing instances of certain grammatical constructions, in order to test questions about language acquisition (Patil et al., 2024; Misra and Mahowald, 2024). Following this line of reasoning, in this paper, we train language models to have specific L1 and L2 language experience, which would be extremely difficult if not impossible to do with human participants.

Small Models and Syntactic Learning Syntax is learned very early in training by language models (Blevins et al., 2022; Chang et al., 2024b). Even models trained on human-scaled training data quantities (around 100M words; Warstadt et al., 2023) show robust syntactic generalizations. Small scale models also permit more manipulations of the training data, given a fixed compute budget, and lend themselves to interpretability analyses.

Training small, controlled models exemplifies the "controlled rearing" (Misra and Mahowald, 2024) approach, in which models are carefully trained with respect to their data exposure in order to make inferences about the effect of training data on model learning. This is only possible for many researchers at a smaller scale.

Bilingual Models The bilingual models trained in this paper resemble those in other recent studies using controlled bilingual models to investigate linguistically motivated questions. Aoyama and Schneider (2024) first train bilingual models on a first language (L1), freeze some model parameters, then continue training with data from the second language (L2). Constantinescu et al. (2025) train bilingual models with different conditions, similar to our "interleaved" and "simultaneous" bilingual conditions.

3 Training Bilingual Language Models

We pre-train bilingual language models from scratch to simulate the language experience of bilingual participants in human crosslingual structural priming experiments. We have two conditions. In the **simultaneous bilingual** condition, the models are exposed only to L1 during the first half of training, then an equal mix of L1 and L2 data in the second half. Models in the **sequential bilingual** condition are exposed only to L1 during the first half of training, then only to L2 in the second half.

We manipulate three factors: language pair (English-Dutch, English-Spanish, English-Polish, English-Greek), language exposure order (e.g. English L1, Dutch L2 vs. Dutch L1, English L2), and bilingual condition (simultaneous or sequential). This results in a total of 16 language models. For example, we train 4 Dutch models: Dutch-English simultaneous, Dutch-English sequential, English-Dutch simultaneous, and English-Dutch sequential.

Each model is an autoregressive GPT-2 Transformer language model with 124M parameters (Radford et al., 2018, 2019). Following Chang et al. (2024b), for each language, we take the first 128M lines of the deduplicated OSCAR corpus (Abadji et al., 2021). We train a separate SentencePiece tokenizer (Kudo and Richardson, 2018) for each model, using the same language proportions as the model training data.[1] Each sequence is monolingual, but in mixed conditions, batches have alternating L1 and L2 sequences. We create sequences of 128 tokens, shuffle the sequences, and sample 2B

---

[1] For the simultaneous bilingual condition, the overall training data the model sees is 75% L1 and 25% L2 data. For the sequential bilingual condition, the overall proportions are 50% L1 and 50% L2 data.

tokens for the training set per language (along with 1M tokens per language for evaluation). In total, each model is trained for 128,000 steps. Starting at step 64,000, each model is trained on either a mix of L1 and L2 (simultaneous condition) or only L2 data (sequential condition). We save checkpoints at regular intervals during training, with increased density just after the introduction of L2, halfway through training. Training details are reported in Appendix C. We call these the **B-GPT models**, and release all checkpoints on Hugging Face.

3.1 Loss Patterns

For each checkpoint, we report the mean surprisal (i.e. log-perplexity or eval loss) on the held out evaluation dataset for both languages each model is trained on (Figure 2). In the simultaneous bilingual condition, we observe consistent patterns: L1 mean surprisal goes down quickly in the first half of training, while L2 mean surprisal stays relatively high. After the introduction of L2 at the halfway point, L2 loss drops quickly. Loss for both languages continues to slowly fall for the rest of training. The patterns are quite different for the sequential condition models in the second half of training. After the switch from training on L1 to L2 data, the mean surprisal for the L1 rises sharply. Mean surprisal stays high for the rest of training. This is consistent with catastrophic forgetting (McCloskey and Cohen, 1989), reflecting the drastic shift in the distribution of training text from L1 to L2.

While all models show similar patterns, the relative mean surprisals do differ somewhat across language pairs. For the simultaneous models, especially when English is the L2, there seems to be a language similarity effect. In the second column in Figure 2, by the end of training, there is a much smaller difference between mean surprisal for English and Dutch and English and Spanish, relative to the differences in mean surprisal between English and Polish and English and Greek. The lower the mean surprisal for English, the greater the transfer benefit is from the L1. In the case of Dutch, which is the most similar to English of the four languages, the English performance benefits the most. For the Greek-English model, which is typologically and orthographically distinct from English, the English performance gets less of a boost. This is consistent with work showing that linguistic similarity is one of the best predictors for successful crosslingual transfer (Chang et al., 2024a).

In the sequential condition, especially when English is the L1 (Figure 2, second column from the right), there are differences in the magnitude of the catastrophic forgetting effect. For Dutch the increase in English mean surprisal is less than the increase for the Spanish and Polish, which in turn is less than that for Greek. This also may be due to differences in linguistic similarity (§5.2).

4 Structural Priming Effects

We detect structural priming effects by comparing the relative likelihood of a target sentence after different prime sentences, usually pairs of sentences that are semantically identical but vary in their syntax. These are referred to as *grammatical alternations*. If a language model assigns a higher probability to a sentence after a sentence of the same grammatical structure than after a sentence of a different structure, then we consider the language model to exhibit structural priming effects.

4.1 Calculating Structural Priming Effects

Following analyses in human studies, structural priming effects are computed as the difference in normalized probability of a target sentence following each prime. For example, we compute the normalized probability P_N of a PO target T_{PO} following a PO prime P_{PO} as shown below, where T_{DO} is the DO target and P_{DO} would be a DO prime.

$$P_N(T_{PO}|P_{PO}) = \frac{P(T_{PO}|P_{PO})}{P(T_{PO}|P_{PO}) + P(T_{DO}|P_{PO})}$$

To test for a structural priming effect, we compare $P_N(T_{PO}|P_{PO})$ and $P_N(T_{PO}|P_{DO})$. If the former is significantly higher, i.e. the target following a matching or congruent prime has a higher probability, this would indicate structural priming. For each model and language combination, we fit a linear mixed effects model predicting the normalized probability of the target with prime type as a fixed effect and experimental item as a random intercept. Here, we only report results for the final model checkpoint, but we conduct the same tests for each model checkpoint. We report the results for the other checkpoints in §4.4. After fitting each linear mixed effects model, we correct for multiple comparisons by controlling for false discovery rate (Benjamini and Hochberg, 1995).

4.2 Experimental Materials

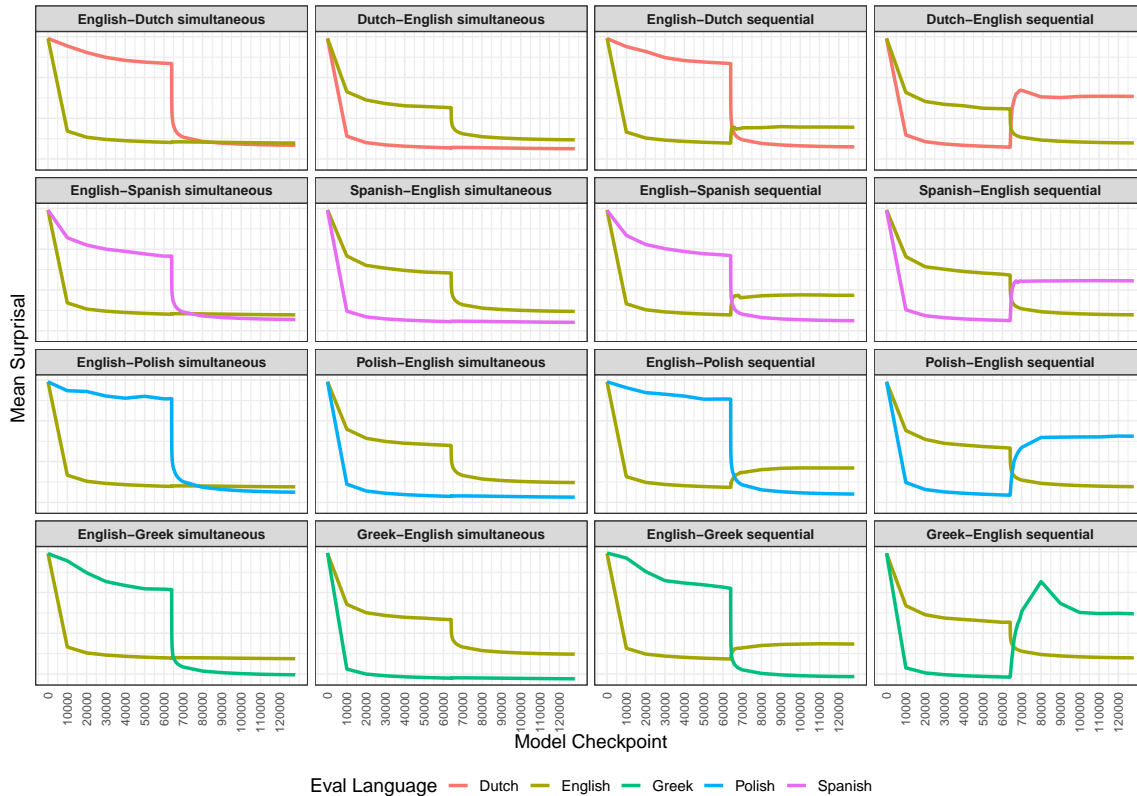We use the experimental stimuli from five studies across the four language pairs, covering three

Figure 2: L1 and L2 mean surprisal for all models and all checkpoints. The color of each line indicates the evaluation language. Each facet represents one model.

grammatical alternations: DO/PO, s-genitive/of-genitive, and Active/Passive (Schoonbaert et al., 2007; Bernolet et al., 2013; Hartsuiker et al., 2004; Fleischer et al., 2012; Kotzochampou and Chondrogianni, 2022). We provide further descriptions and examples in Appendix A. We check whether the items appear in the training data and report results in Appendix B.

DO/PO The first alternation is for ditransitive events, i.e. events with two objects. One of these is the Prepositional Object (PO) construction (Ex. (1)). In this construction, the direct object 'hat' directly follows the verb and the indirect object is introduced with a prepositional phrase 'to the boxer'. The other is the Double Object (DO) construction (Ex. (2)). In this construction, the indirect object 'boxer' follows the verb, followed immediately by the direct object 'hat'. Dutch has a comparable alternation.

(1)   The cook shows a hat to the boxer.   (PO)

(2)   The cook shows the boxer a hat.   (DO)
        (Schoonbaert et al., 2007)

s-genitive/*of*-genitive The second alternation is for genitive constructions, which encode information about possession. In English, one of the constructions is the s-genitive (Ex. (3)), where the possessor 'nun' is marked with ''s' and the possessor 'nun' precedes the possessed thing 'egg'. In the of-genitive construction (Ex. (4)), the order is reversed and the preposition 'of' is used to express the possessive relationship. Dutch also has this alternation.

(3)   The nun's egg is yellow.   (s-gen)

(4)   The egg of the nun is yellow.   (of-gen)
        (Bernolet et al., 2013)

Active/Passive Finally, many events can be encoded using either active or passive voice. In active sentences like Ex. (5), the agent, or do-er of the action, in this case 'the taxi', is the syntactic subject of the sentence. The theme or patient, i.e. the thing having an action done to it, 'truck' in this case, is the syntactic object of the sentence and follows the noun. In passive sentences, the syntactic subject of the sentence is the theme and the agent is introduced in a prepositional phrase, 'by the taxi' (Ex. (6)).
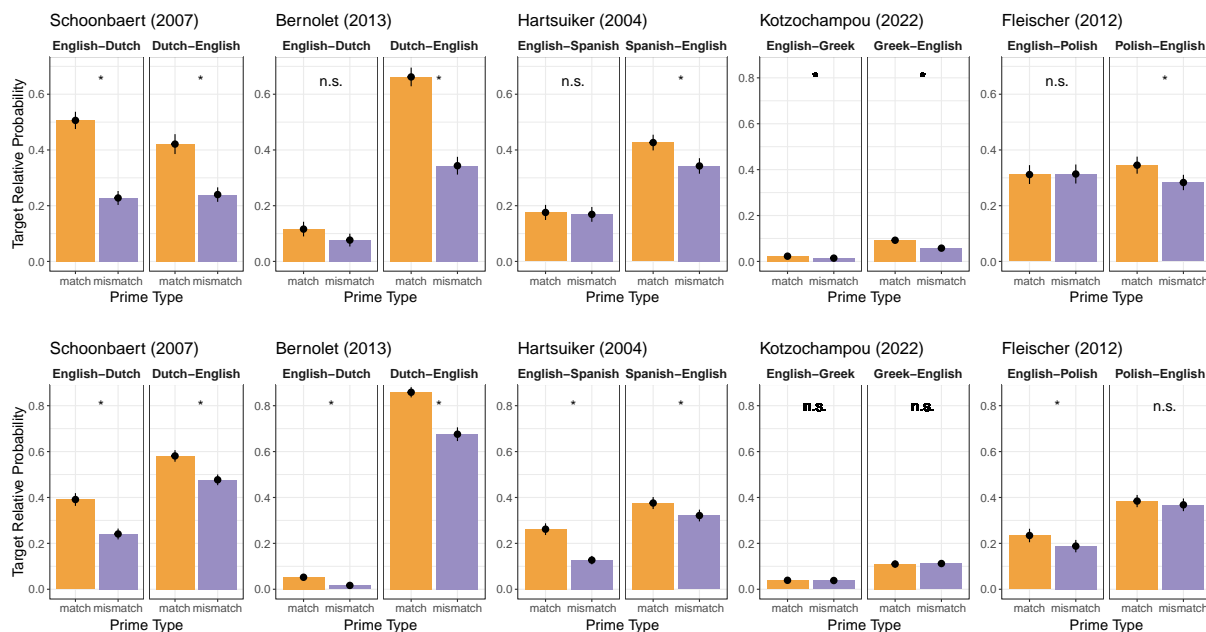
Figure 3: Priming results for the simultaneous (top) and sequential (bottom) bilingual models. For all experiments, prime language corresponds to L1 and target language corresponds to L2. Significance is indicated with *. Color indicates prime condition. Orange indicates congruent or matching prime and target types and purple indicates mismatched prime and target types. Specific grammatical alternations tested are described in Appendix A.

(5)   The taxi chases the truck.   (Active)

(6)   The truck is chased by the taxi.   (Passive)
        (Hartsuiker et al., 2004)

For each alternation, there are two grammatical constructions which convey the same information and differ primarily in their syntax. For each language pair, both languages share the same grammatical alternation. For example, English and Spanish both share the active/passive alternation. Therefore, for example, we test whether English actives prime Spanish actives and vice versa.

The original Spanish, Greek, and Polish experiments have many fewer stimuli pairs than the Dutch experiments. Because we do not primarily aim to replicate human experimental results, we create new prime-target pairs by considering every possible pair of prime and target sentences. Then, we randomly sample pairs so that we have 144 pairs each for the Spanish, Greek, and Polish stimuli. This matches the amount of statistical power for the Dutch experimental materials.

4.3   Results

Overall, we replicate the crosslinguistic structural priming effects[2] in Michaelov et al. (2023) (Fig-

ure 3, top). In all cases, when English is the target language, we find that a target sentence is more likely if the prime sentence matches its grammatical structure. We also find statistically significant structural priming effects for the experiments with Schoonbaert et al. (2007) and Kotzochampou and Chondrogianni (2022) stimuli when English is the prime language. There is still a numerical effect in the expected direction for the experiments with Bernolet et al. (2013) and Hartsuiker et al. (2004) stimuli where English is the prime language.

However, there remains an asymmetry in the results, where we see more robust structural priming effects when English is the target language, as opposed to when English is the prime language. We discuss this in depth in Section 5.1.

Notably, we also find structural priming effects in the final checkpoints of the sequential bilingual models (Figure 3, bottom) for some languages, despite evidence that the models experienced catastrophic forgetting of L1 (§3.1). All of the Dutch and Spanish models still exhibit structural priming effects in the final checkpoints, and we see significant structural priming in the English-Polish model. However, there is a reduced effect size,

---

[2]Following results from the human structural priming literature, where it has been found that structural priming effects

are strongest when the prime language is the participant's L1, and the target language is the L2, we only report results from the L1→L2 priming conditions. We report L2-L1 priming results in Appendix D.
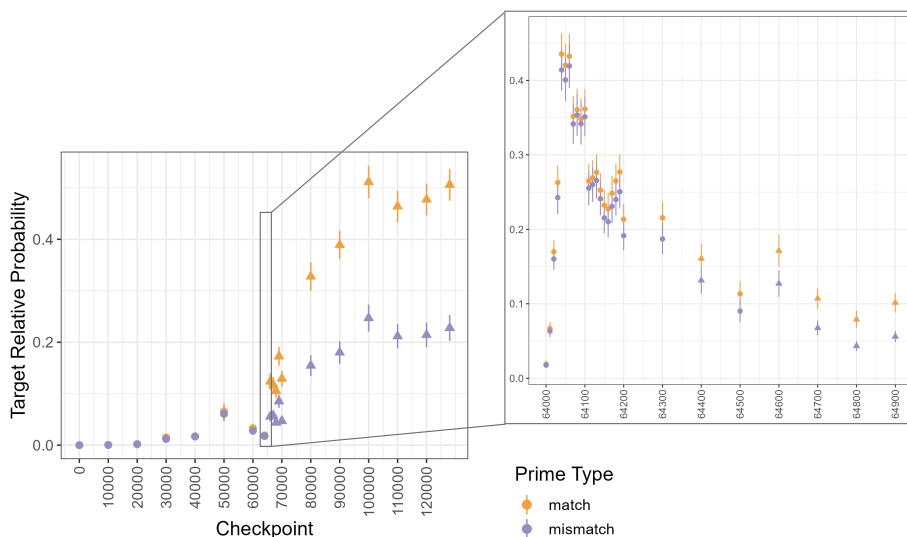
Figure 4: The panel on the left shows structural priming effects for English-Dutch priming for the simultaneous bilingual model, evaluated on Schoonbaert et al. (2007) stimuli. Significant structural priming effects are marked with triangles, and effects that are not significant are marked with circles. In the panel on the right, we plot the structural priming effects for the first 900 steps after L2 exposure.

likely caused by the catastrophic forgetting, where L1 knowledge is less well-represented by the end of training despite the fact that shared grammatical representations remain present to some degree. The stronger effects for Dutch and Spanish, and less strong effects for Greek and Polish, are likely an effect of language similarity with English (§5.2).

4.4 Training Dynamics

Next, we characterize the time course of the models' learning of shared representations. We first check that structural priming effects are temporally linked to L2 proficiency, because if the models demonstrate structural priming effects before being exposed to L2, we can infer that structural priming is possible through exposure to L1 alone (e.g. due to data contamination across languages).

To test this, we use BLiMP (Warstadt et al., 2020) to measure L2 proficiency at each checkpoint.[3] BLiMP measures the grammatical knowledge of the model, which is predictive of a model's ability to generate grammatical text. We evaluate each model checkpoint on BLiMP using the LM Evaluation Harness (Biderman et al., 2024), and we report the average score over all tasks. We report results for all models in Appendices E and F.

We then evaluate structural priming at each

model checkpoint (e.g. Figure 4 for the English-Dutch simultaneous bilingual model). Before the model is exposed to L2 data, there are no priming effects. But shortly after exposure to L2—as early as 600 steps after exposure to L2, or 4.9M L2 tokens—the language model exhibits stable priming effects. We then compare the time course of structural priming effects to language proficiency. Figure 5 show structural priming effects as the difference in the relative probabilities between the matching and mismatching prime, plotted in black. In pink, we show the English BLiMP scores.

In the simultaneous bilingual condition (Figure 5, top), structural priming effects emerge at the same time as the model shows a jump in BLiMP performance. Therefore, we argue this draws a stronger link between structural priming behavior and shared multilingual representations. In the sequential bilingual condition, we plot L2 English BLiMP accuracy. In the second half of training, accuracy drops as a result of catastrophic forgetting, but structural priming effects still appear and stay relatively high over the course of training. Therefore, it seems that even when the model experiences catastrophic forgetting, representations may still be shared between languages and allow for transfer learning. However, this effect is most clear for Dutch, which is most similar to English. For the other languages, especially Polish and Greek, structural priming effects do not persist after catas-

---

[3]While there are BLiMP benchmarks for other languages, BLiMP does not exist for all other languages in our sample. Therefore, we limit our analysis to English BLiMP.
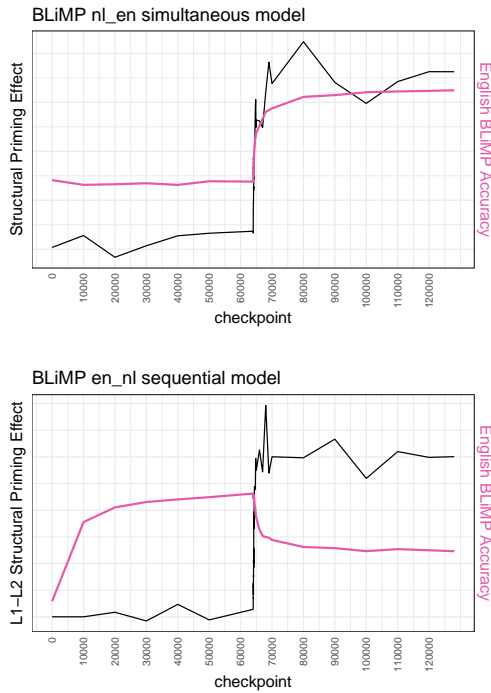
Figure 5: Structural priming effect and BLiMP accuracy over training for Dutch-English simultaneous (top) and English-Dutch sequential (bottom) models.

trophic forgetting. This is likely another language similarity effect (§5.2). We report comparisons of priming effects and BLiMP accuracy for all models in Appendix E.

5 Discussion

5.1 Language Asymmetries

In human structural priming experiments, it has been shown that structural priming effects are generally stronger in L1→L2 priming (e.g. Schoonbaert et al., 2007), although in some language pairs, there are no priming effects at all. Shin and Christianson (2009) showed evidence of Korean-English priming, but Shin and Christianson (2011) found no English-Korean structural priming effects. These experiments have a serious confound, however, as participants are always L2 English speakers. Therefore it is not possible to determine through these experiments whether effect asymmetries are due to L1→L2 versus L2→L1 priming or due to the target language being English. In this paper, we found that there were stronger priming effects when English was the target language, independent of its L1/L2 status and when controlling for language exposure. Therefore we argue that the results in the psycholinguistics literature may not be due to differences in L1→L2 and

L2→L1 priming, but may be driven by whether English is the target language.

The experiments in this paper rule out the role of model training data quantity, which suggests the asymmetry may be due to cross-linguistic differences. It is possible that there is something about English as a target language that increases structural priming effects. One candidate is sensitivity to word order. In contrast to English, Polish and Greek are morphologically rich languages, where important information is conveyed through morphology (e.g. word inflections), and word orders are less fixed (Tzanidaki, 1995; Siewierska, 1993). Polish and Greek showed less robust structural priming effects across all conditions relative to Dutch and Spanish. Similarly, in human experiments, there is a demonstrated asymmetry for Korean, which also has overt morphological marking and less fixed word order. In Tagalog, a language with even more flexible word order, there is evidence from within-language priming that Tagalog speakers do not exhibit structural priming effects based on word order (Garcia and Kidd, 2020; Garcia et al., 2023). Therefore, taken together with work in psycholinguistics, the results in this paper call for a reconsideration of the interpretation of previous experimental work. The asymmetries in structural priming effects may be attributed to crosslinguistic differences in the importance of word order, rather than L1/L2 status.

This result serves as an example of the value of language models as model organisms. Disentangling the role of L1→L2 priming and the role of English as target language is difficult to do with human participants, because it is much easier to find participants for whom English is an L2 than English L1 speakers who speak another language to a high level of proficiency. Our experiments demonstrate the value of language model experiments to develop and refine hypotheses in psycholinguistics that can then be validated through human studies.

5.2 Language Similarity

In the experiments presented above, there were effects of language similarity throughout. There is a marked difference between the robustness of structural priming effects for Polish and Greek, relative to Dutch and Spanish. In the sequential bilingual condition, the structural priming effects are more robust to catastrophic forgetting when the language pairs are more closely related. In these cases, when we see evidence of catastrophic forgetting, structural priming effects are still present for Dutch and

Spanish, but not Greek and Polish. This suggests that in the case of catastrophic forgetting, language similarity is a key factor in the extent to which existing L1 representations will persist after a significant distributional shift in the training data.

As our sample of languages is small and comes from one language family, it is not possible to quantitatively analyze the impact of various typological features. Instead, we explore some possible relevant differences that may affect structural priming effects, such as writing system and how grammatical alternations are encoded. We provide examples of the alternations and further description in Appendix A. English, Dutch, Spanish, and Polish all use periphrastic constructions to encode the passive voice, whereas Greek uses verbal morphology to do so. In English, the difference between the active and passive verb forms is seen in (5) and (6), where the passive is a periphrastic form where the present form of the verb 'to be' is combined with the past participle of 'chase'. By contrast, Greek has a specific verbal morphology to encode active or passive voice. This is unlike the other languages included in our experiments, which use a combination of the present copula and the past participle to mark passive voice.

Both of these are typological differences. With respect to orthography, Greek is the only language in this set of experiments that uses a non-Latin writing system. Therefore, there is essentially no vocabulary overlap between English and Greek, while the other language pairs may have tokens shared between the languages. Compounding with typological differences, this differing orthography and lack of shared tokens may contribute to the reduced structural priming effects observed between English and Greek.

By studying shared multilingual representations in language models, our results also tie to work in crosslingual transfer in language models. Chang et al. (2024a) show that language relatedness—especially syntactic typological similarity—is predictive of how much benefit there is to adding multilingual data to improve performance for a target language, relative to a monolingual setting. Thus, our results are consistent with previous work showing that crosslingual transfer is more effective between more similar languages. This not only provides a better understanding of crosslingual transfer, but it is indicative of the general limitations of crosslingual transfer. Even for languages in the same language family (in this case, Indo-European), there is

still limited ability for models to successfully create shared abstract grammatical representations for language pairs such as Greek and English, relative to a closely related language pair like Dutch and English. Therefore, we argue that these results suggest the reconsideration of some current practices for leveraging crosslingual transfer. A common approach for developing a model, especially for a low-resource language, is to start with a powerful open-weight model primarily trained on English and do continued pre-training, vocabulary adaptation, etc. to improve performance for the target language. Our results support previous work showing that using models trained on less data from more similar languages leads to competitive or better results (e.g. Ogueji et al., 2021).

6 Conclusion

In this paper, we used structural priming to understand the shared multilingual representations that drive crosslingual transfer. First, we trained controlled, comparable bilingual language models and replicated crosslingual structural priming effects from previous work. We release the models in order to enable continued work on related questions. We then described the time course of the emergence of structural priming effects relative to the acquisition of L2, drawing a temporal link between L2 proficiency and structural priming effects. We also demonstrated that structural priming effects may persist despite catastrophic forgetting of L1, depending on language similarity between L1 and L2. We argue that language similarity affects several components of this work and should be considered more when attempting to leverage crosslingual transfer in language model development.

Perhaps most notably, the results in this paper show an asymmetry, where priming effects are stronger when English is the target language. We overcome a confound in prior psycholinguistic research and argue that these results suggest a new interpretation of previous results.

Limitations

Language Sample All of the languages we use in the experiments are Indo-European. While we do cover four distinct sub-branches of the Indo-European language family, this language sample is not sufficiently diverse too draw strong, generalizable conclusions. The language sample is primarily driven by the availability of psycholinguistic

datasets, which are more often representative of European languages.

Model Size   The models we train are very small. This is due to compute limitations. If we trained larger models, we may not have seen the same limits on shared representations and crosslingual transfer, as the models would have not reached capacity limitations as easily. In future follow-up work, increasing the model size would likely be necessary in order to study successful crosslingual transfer in language pairs that are more different than English and Greek or English and Polish. Training larger models and how these effects change with model and data scale would also be illuminating, but is currently not possible given our resources.

Our view is that it is best to first establish evidence for a phenomenon with small models. Now that there is evidence of this phenomenon, we can test larger models in the future to test whether these results change as a function of model scale. And, in this case we also aimed to manipulate several factors. Given our limited compute budget and the fact that we were training models from scratch, we would not have been able to do as many manipulations if we had trained bigger models. Additionally, smaller models more easily allow mechanistic interpretability work, so we feel these models are more useful and accessible at this scale.

Language Data Contamination   While we argue that asymmetries in structural priming effects are due to language differences, it is also possible that the asymmetries could be due to data contamination. If the non-English data could be contaminated with English data, in the cases where English is the target language, the model would see more English data than intended because of contamination. This could boost the structural priming effects, especially when English is the target language.

Similarly, in Figure 2, there is an asymmetry between the English-Dutch and Dutch-English simultaneous models, where the English L2 loss drops much more quickly in the first half of training than does the loss for Dutch as L2. When Dutch is the L1, the model is supposedly not being trained on English. We hypothesize that this is due to English contamination in the Dutch data. The reason we see an asymmetry is likely because there is not as as much Dutch contamination in the English data. This could be due to language use: many Dutch people speak English, but proportionally not as many English speakers also speak Dutch. It could also be

due to differences in accuracy of language identification (LID) methods for English and Dutch, as English and Dutch are highly similar languages.

Ethical Considerations

We do not believe the work in this paper raises ethical concerns, but instead we hope it contributes to a better understanding of multilingual language models and indirectly making language models better for more languages.

Acknowledgements

---

[4]Carbon emissions were calculated via `https://mlco2.github.io/impact/#compute`.

[5]`https://osf.io/5cw2e/`

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021*, pages 1–9, Mannheim. Leibniz-Institut für Deutsche Sprache.

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.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Sarah Bernolet, Robert J. Hartsuiker, and Martin J. Pickering. 2013. From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition*, 127(3):287–306.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the Trenches on Reproducible Evaluation of Language Models. *arXiv preprint arXiv:2405.14782*.

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

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024a. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2024b. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *Transactions of the Association for Computational Linguistics*, 12:1346–1362.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7570–7577.

Sunjoo Choi and Myung-Kwan Park. 2022. Syntactic priming in the L2 neural language model. *The Journal of Linguistic Science*, 103:81–104.

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.

Zuzanna Fleischer, Martin J. Pickering, and Janet F. McLean. 2012. Shared Information Structure: Evidence from Cross-Linguistic Priming. *Bilingualism: Language and Cognition*, 15(3):568–579.

Stefan Frank. 2021. Cross-language structural priming in recurrent neural network language models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

Richard Futrell and Kyle Mahowald. 2025. How Linguistics Learned to Stop Worrying and Love the Language Models. *arXiv preprint arXiv:2501.17047*.

Rowena Garcia and Evan Kidd. 2020. The acquisition of the tagalog symmetrical voice system: Evidence from structural priming. *Language Learning and Development*, 16(4):399–425.

Rowena Garcia, Jens Roeser, and Evan Kidd. 2023. Finding your voice: Voice-specific effects in tagalog reveal the limits of word order priming. *Cognition*, 236:105424.

Robert J. Hartsuiker, Martin J. Pickering, and Eline Veltkamp. 2004. Is Syntax Separate or Shared Between Languages?: Cross-Linguistic Syntactic Priming in Spanish-English Bilinguals. *Psychological Science*, 15(6):409–414.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Shailee Jain, Vy A Vo, Leila Wehbe, and Alexander G Huth. 2024. Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, 5(1):80–106.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14727–14742, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sotiria Kotzochampou and Vasiliki Chondrogianni. 2022. How similar are shared syntactic representations? Evidence from priming of passives in Greek–English bilinguals. *Bilingualism: Language and Cognition*, 25(5):726–738.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

James Michaelov, Catherine Arnett, Tyler Chang, and Ben Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

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.

Stefan Müller. 2024. Large language models: The best linguistic theory, a wrong linguistic theory, or no linguistic theory at all? *Lingbuzz Preprint*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Steven T Piantadosi. 2023. Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414.

Martin J Pickering and Victor S Ferreira. 2008. Structural priming: a critical review. *Psychological bulletin*, 134(3):427.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

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

Sofie Schoonbaert, Robert J. Hartsuiker, and Martin J. Pickering. 2007. The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language*, 56(2):153–171.

Jeong-Ah Shin and Kiel Christianson. 2009. Syntactic processing in Korean–English bilingual production: Evidence from cross-linguistic structural priming. *Cognition*, 112(1):175–180.

Jeong-Ah Shin and Kiel Christianson. 2011. The Status of Dative Constructions in Korean, English and in the Korean-English Bilingual Mind. *Processing and producing head-final structures*, pages 153–169.

Anna Siewierska. 1993. Syntactic weight vs information structure and word order variation in Polish. *Journal of Linguistics*, 29(2):233–265.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Dimitra Irini Tzanidaki. 1995. Greek word order: towards a new approach. *UCL Working Paper in Linguistics*, 7:247–277.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

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.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Crosslingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Zhenghao Zhou, Robert Frank, and R Thomas McCoy. 2024. Is in-context learning a type of gradient-based learning? evidence from the inverse frequency effect in structural priming. *arXiv preprint arXiv:2406.18501*.

A Grammatical Alternations

DO/PO We use the Dutch and English stimuli from Schoonbaert et al. (2007), which contain pairs that contrast the Prepositional Object (PO) and Double Object (DO) dative constructions.

In some languages, for ditransitive sentences, when there are two objects, there are two possible ways to express the same event. One of these is the **Prepositional Object (PO)** construction (Ex. (7-a)). In this construction, the direct object 'hat' directly follows the verb and the indirect object is introduced with a prepositional phrase 'to the boxer'. The other is the **Double Object (DO)** construction (Ex. (7-b)). In this construction, the indirect object 'boxer' follows the verb, followed immediately by the direct object 'hat'.

(7) a. The cook shows a hat to the boxer. (PO)
     b. The cook shows the boxer a hat. (DO)

     (Schoonbaert et al., 2007)

Dutch has an equivalent alternation, with the same word order as English for PO (Ex. (8-a)) and DO (Ex. (8-b)) sentences.

(8) a. De kok toont een hoed aan de bokser.
         The cook shows a hat to the boxer.
     b. De kok toont de bokser een hoed.
         The cook shows the boxer a hat.
         (Schoonbaert et al., 2007)

***s*-genitive/*of*-genitive** We use the Dutch and English stimuli from Bernolet et al. (2013), which contrast the two genitive constructions, which are semantically equivalent ways to express possession. In English, one of these is the **s-genitive** construction (Ex. (9-a)), where the possessor 'nun' is

marked with ''s'. In this construction, the possessor 'nun' precedes the possessed thing 'egg'. In the **of-genitive** construction (Ex. (9-b)), the order is reversed and the possessed thing precedes the possessor. In this case, the preposition 'of' is used to express the possessive relationship.

(9)   a.   The nun's egg is yellow.   (s-gen)
       b.   The egg of the nun is yellow. (of-gen)
         (Bernolet et al., 2013)

Dutch has a similar alternation. For proper names, s-genitive possession can be marked with ''s', but for common nouns, possession is marked with the possessive pronoun that corresponds in gender to the possessor noun. In the example below (Ex. (10-a)), *non* 'nun' is feminine, so *haar* 'her' marks possession. Masculine nouns use *zijn* 'his' (Bernolet et al., 2013). The Dutch of-genitive construction is more similar to English, where the preposition *van* 'of' is used to show possession, and the order of the possessor and possessee is flipped, relative to the s-genitive order.

(10)   a.   De non haar   ei   is geel.
         The nun POSS egg is yellow.
       b.   Het ei   van de   non is geel.
         The egg of   the nun is yellow.
         (Bernolet et al., 2013)

Active/Passive For Spanish-English, Polish-English, and Greek-English experiments, we use stimuli that contrast active and passive constructions. For Spanish-English, we use stimuli from Hartsuiker et al. (2004); for Greek-English, the stimuli come from Kotzochampou and Chondrogianni (2022); and for Polish-English, we use stimuli from Fleischer et al. (2012).

Many languages allow events to be expressed as either active or passive. In **active** sentences, e.g. Ex. (11-a), the agent, or do-er of the action, 'the taxi' is the syntactic subject of the sentence, which in English, is marked by being the first argument in the sentence. The theme or patient, i.e. the thing having an action done to it, 'truck' is the syntactic object of the sentence and follows the noun. In **passive** sentences, the syntactic subject of the sentence is the theme. The agent is introduced in a prepositional phrase, 'by the taxi' (Ex. (11-b)).

(11)   a.   The taxi chases the truck.   (Active)
       b.   The truck is chased by the taxi. (Passive)
         (Hartsuiker et al., 2004)

Spanish expresses active and passive sentences very similarly to English, following the same word order (Ex. (12-a) and (12-b), respectively).

(12)   a.   El   taxi persigue el   camión.
         The taxi chases   the truck.
       b.   El   camión es perseguido por el   taxi.
         The truck   is chased   by   the taxi.
         (Hartsuiker et al., 2004)

Typologically, Polish and Greek are more different from English than either Dutch or Spanish is. Both of these languages mark the syntactic subjects and objects using case marking, unlike English, Dutch, and Spanish, which do this only with word order. In Polish, for example, in the active, *sportowiec* 'sportsman' is in the nominative case and is the syntactic subject of the sentence. The patient 'ballet dancer' takes the accusative and is the grammatical object of the sentence. In the passive, it is in the accusative case (*sportowca*) and is introduced with a prepositional phrase. The patient 'ballet dancer', in this case, is in the nominative case.

(13)   a.   Sportowiec
         sportsman.NOM.SG
         przygniata
         squash.PRES.3SG
         baletnicę.
         ballet-dancer.ACC.SG
         "The sportsman squashes the ballet dancer."
       b.   Baletnica   jest
         ballet-dancer.NOM.SG be.3SG.PRES
         przygniatana   przez
         squash.PST.PART by
         sportowca.
         sportsman.ACC.SG
         "The ballet dancer is squashed by the sportsman."
         (Fleischer et al., 2012)

Similarly, Greek marks subject and object roles with case marking. When it is the subject, αθλητής (*athlitis*) 'athlete' is nominative, but as an object, it takes the accusative case (αθλητή, *athliti*). Greek, unlike Polish or the other languages described here, has a specific verbal morphology to encode active or passive voice (compare (14-a) and (14-b)), therefore the verb form is also specific to passive voice, unlike the other languages shown here, which use a combination of the present copula and the past participle to mark passive voice.

(14)   a.   Ο αθλητής κλωτσάει τον κλέφτη.
         O athlitis   klots**aei**   ton klefti.

The athlete.NOM kicks-ACTIVE the thief.ACC.

"The athlete kicks the thief."

b. Ο κλέφτης κλωτσιέται από τον

O kleftis klots**iete** apo ton

αθλητή.

athliti.

The thief.NOM kicks-**PASSIVE** by the athlete.ACC.

"The thief is kicked by the athlete."

(Kotzochampou and Chondrogianni, 2022)

B Contamination Analysis

The "What's in My Big Data?" tool[6] indexes OSCAR and allows n-gram search, but we were unable to access it. Instead, we use Infini-gram[7] (Liu et al., 2024), which indexes C4, which is also compiled of multilingual web data and is much larger than the portion of OSCAR we used to train our models. Only a very small number of our stimuli can be found in C4 (Table B.1).

	Contaminated Items	
Dataset	**Count**	**Proportion**
Schoonbaert	0	0
Bernolet	0	0
Hartsuiker	3	0.0078
Fleischer	0	0
Kotzochamopou	4	0.0208

Table B.1: C4 contamination results.

C Model Training Details

Model training code is based on that from Chang and Bergen (2022).[8]

Model Hyperparameters Table C.2 shows the model training hyperparameters.

Checkpoints We take checkpoints at the first and last steps (128k). Additionally we take checkpoints every 10k steps. After the introduction of the L2 at the halfway point (64k), we save checkpoints every 10 steps, because we expect that structural priming effects may emerge within the first few hundred training steps after the introduction of L2.

---

[6]https://wimbd.apps.allenai.org/

[7]https://huggingface.co/spaces/liujch1998/infini-gram

[8]Available at https://github.com/tylerachang/word-acquisition-language-models.

After 200 steps after the introduction of L2, we gradually increase the checkpoint intervals. This way, we have increased resolution during the period of training where we expect to see the emergence of structural priming effects, while minimizing the number of checkpoints needed.

We save model checkpoints at the following training steps: 0, 10000, 20000, 30000, 40000, 50000, 60000, 64000, 64010, 64020, 64030, 64040, 64050, 64060, 64070, 64080, 64090, 64100, 64110, 64120, 64130, 64140, 64150, 64160, 64170, 64180, 64190, 64200, 64300, 64400, 64500, 64600, 64700, 64800, 64900, 65000, 66000, 67000, 68000, 69000, 70000, 80000, 90000, 100000, 110000, 120000, 128000.

D L2-L1 Priming

Figures D.6 and D.7 show the L2→L1 results for all models for both the simultaneous and sequential bilingual conditions, respectively. Each facet represents a model. The labels, e.g. English-Dutch and Dutch-English, correspond to the L1 and L2 of each model.

Hyperparameter	**Value**
Layers	12
Embedding size	768
Hidden size	768
Intermediate hidden size	3072
Attention heads	12
Attention head size	64
Activation function	GELU
Vocab size	50004
Max sequence length	128
Position embedding	Absolute
Batch size	128
Train steps	128k
Learning rate decay	Linear
Warmup steps	10000
Learning rate	1e-4
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.999
Dropout	0.1
Attention dropout	0.1

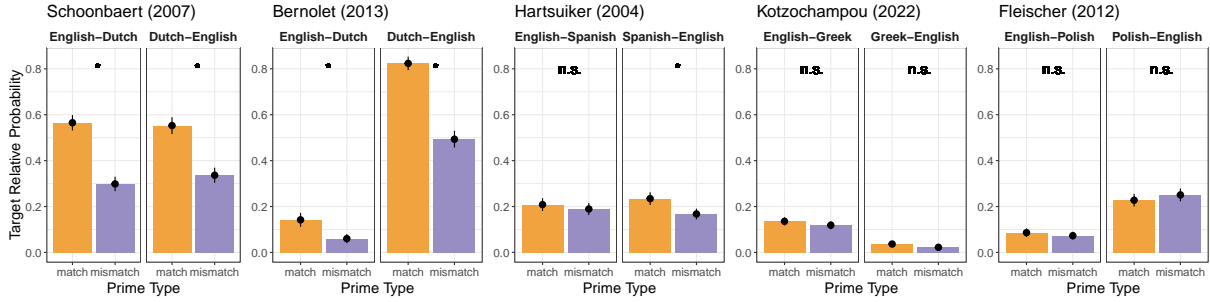Table C.2: Language model hyperparameters.

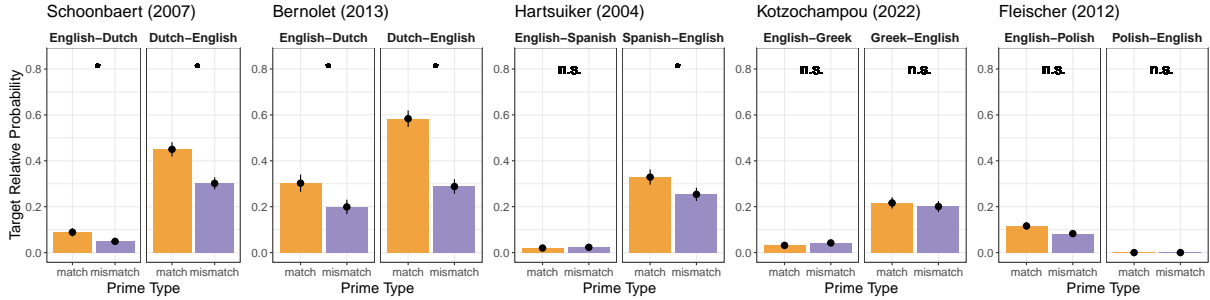Figure D.6: Simultaneous bilingual condition. Prime language corresponds to L2.

Figure D.7: Sequential bilingual condition. Prime language corresponds to L2.

E   Full BLiMP Results

E.1   Schoonbaert (2007)

Figure E.8 shows the comparison for structural priming effects and BLiMP scores for all Dutch-English models.

E.2   Bernolet (2013)

Figure E.9 shows the comparison for structural priming effects and BLiMP scores for all Dutch-English models.

E.3   Hartsuiker (2004)

Figure E.10 shows the comparison for structural priming effects and BLiMP scores for all Spanish-English models.

E.4   Fleischer (2012)

Figure E.11 shows the comparison for structural priming effects and BLiMP scores for all Polish-English models.

E.5   Kotzochampou (2022)

Figure E.12 shows the comparison for structural priming effects and BLiMP scores for all Greek-English models.

F   Supplementary BLiMP Analysis

For models where English is the L1, we see differences in BLiMP scores over the course of training according to the bilingual conditions (Figure F.13). In the simultaneous bilingual condition, there is a small dip in BLiMP score after exposure to L2, but then the scores rise again and stay at ceiling. In the sequential bilingual condition, BLiMP scores fall rapidly after exposure to L2. At about step 80000, performance plateaus. The performance never returns to the level of the model at checkpoint 0, but BLiMP score at the final checkpoint is worse than at checkpoint 10000 for all models. This further supports the observation that the models in the sequential bilingual condition experience catastrophic forgetting. It is even more noteworthy, therefore that the models exhibit structural priming effects during the period where L1 mean surprisal rises and BLiMP scores fall.

Comparing BLiMP performance for the models in the simultaneous condition, we observe a difference in final checkpoint performance. Dutch models have the best performance, followed by Spanish. Greek and Polish again show the worst performance. These results demonstrate differential crosslingual transfer benefits. The language that is the most similar to English (Dutch) leads to the highest BLiMP scores, followed by Spanish, which
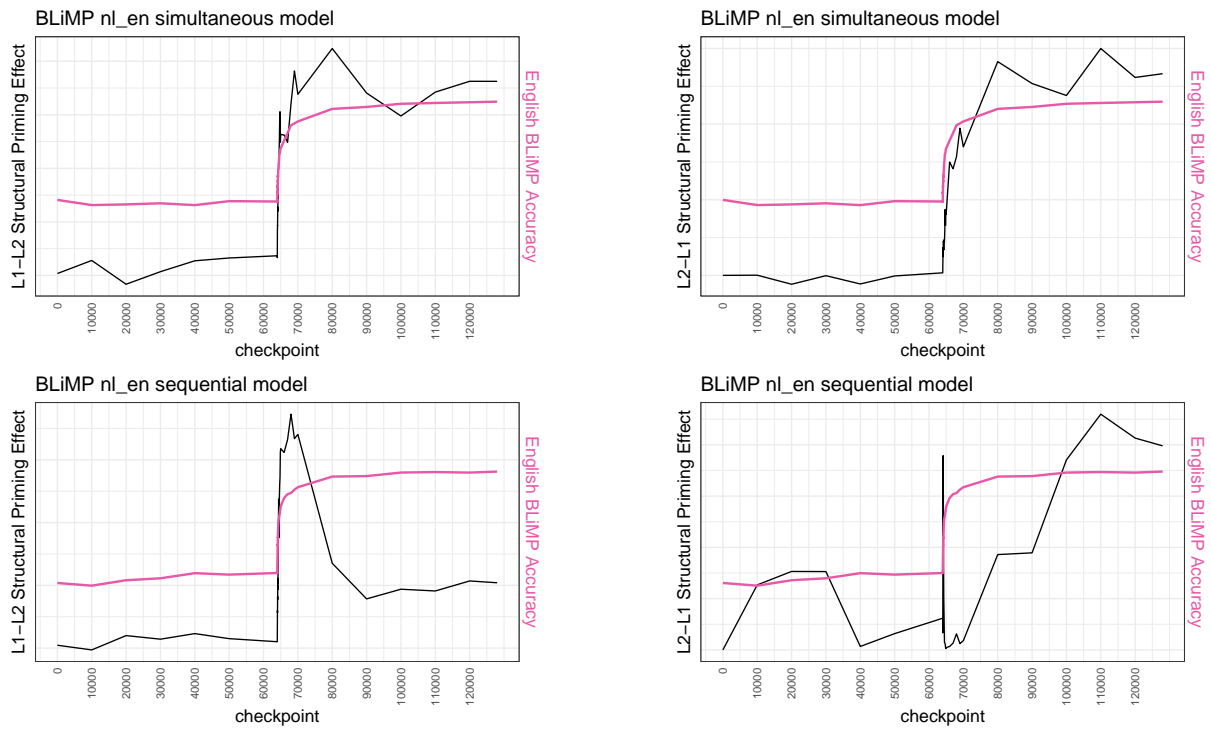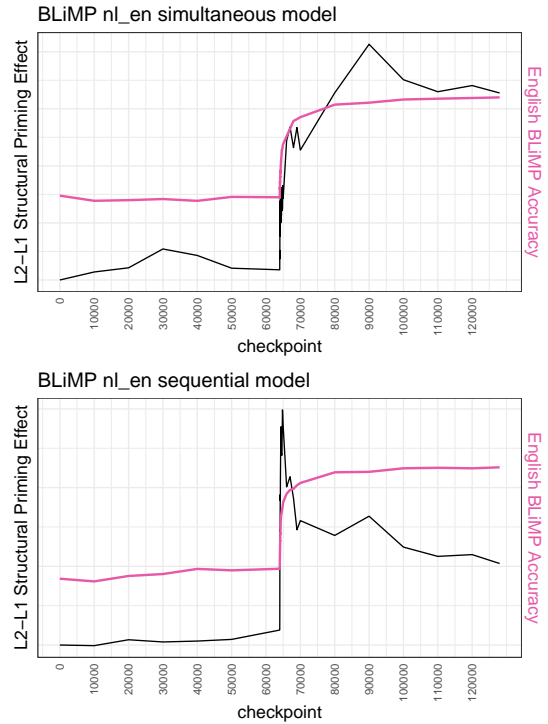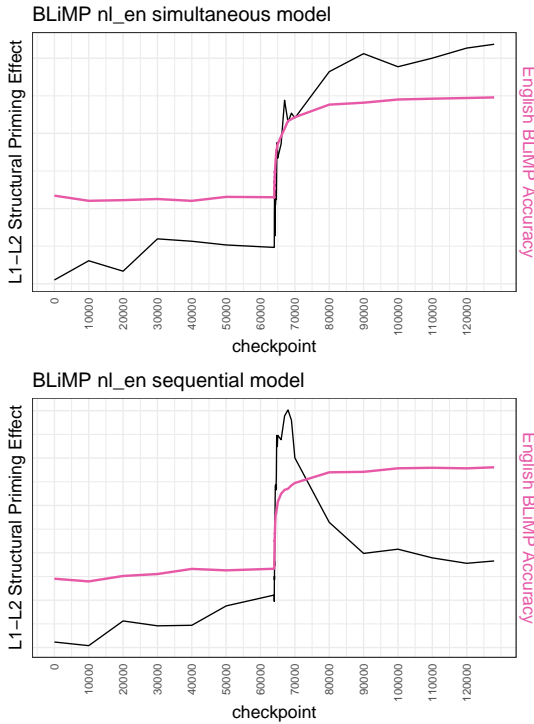
Figure E.8: Structural priming effect (black), plotted as the difference between match and mismatch conditions, and English BLiMP accuracy (pink) over the course of model training. Y-axes have been re-scaled for easier comparison.

is also very similar to English. Polish and Greek are the most different from English and show the least benefit from crosslingual transfer. This is also consistent with previously demonstrated effects of linguistic similarity (Chang et al., 2024a).

Figure E.9: Structural priming effect (black), plotted as the difference between match and mismatch conditions, and English BLiMP accuracy (pink) over the course of model training. Y-axes have been re-scaled for easier comparison.
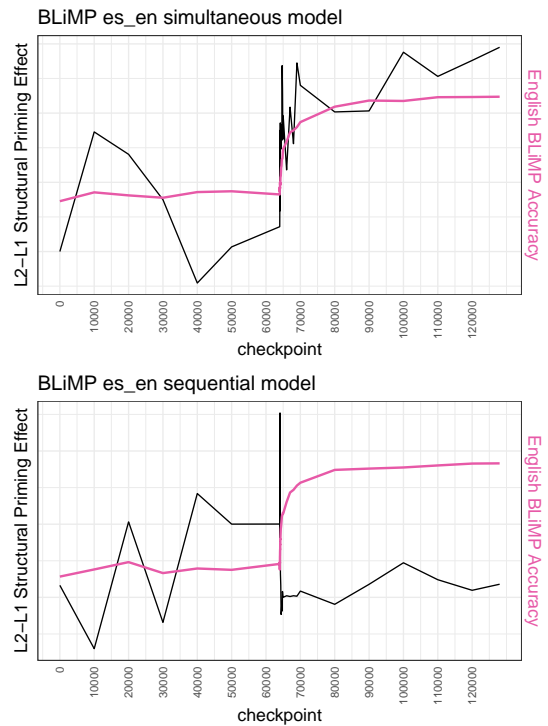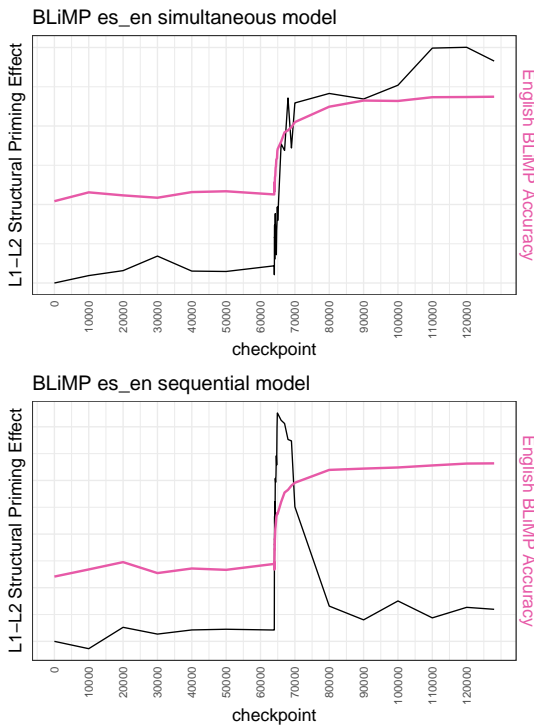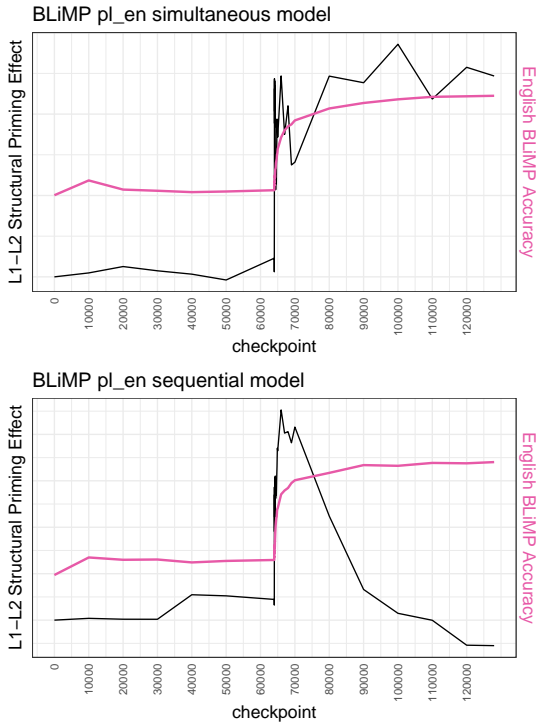

Figure E.10: Structural priming effect (black), plotted as the difference between match and mismatch conditions, and English BLiMP accuracy (pink) over the course of model training. Y-axes have been re-scaled for easier comparison.
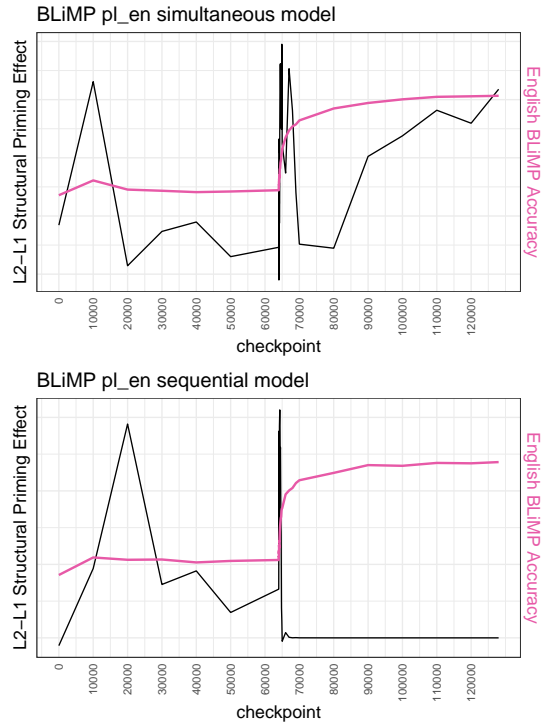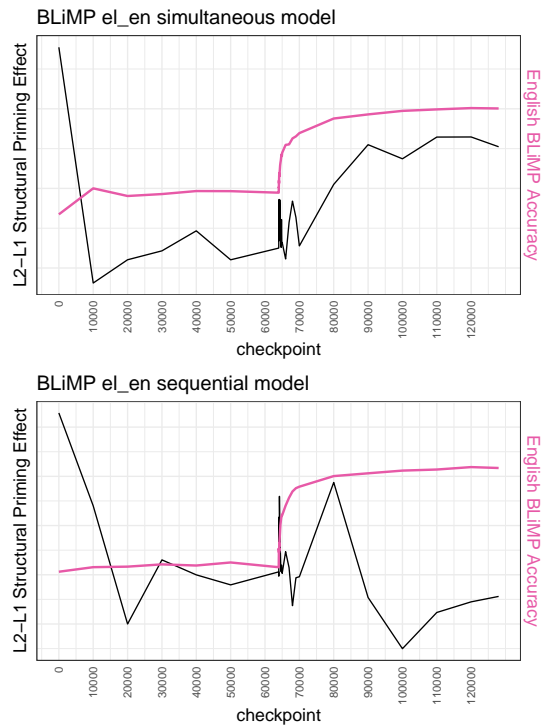
Figure E.11: Structural priming effect (black), plotted as the difference between match and mismatch conditions, and English BLiMP accuracy (pink) over the course of model training. Y-axes have been re-scaled for easier comparison.
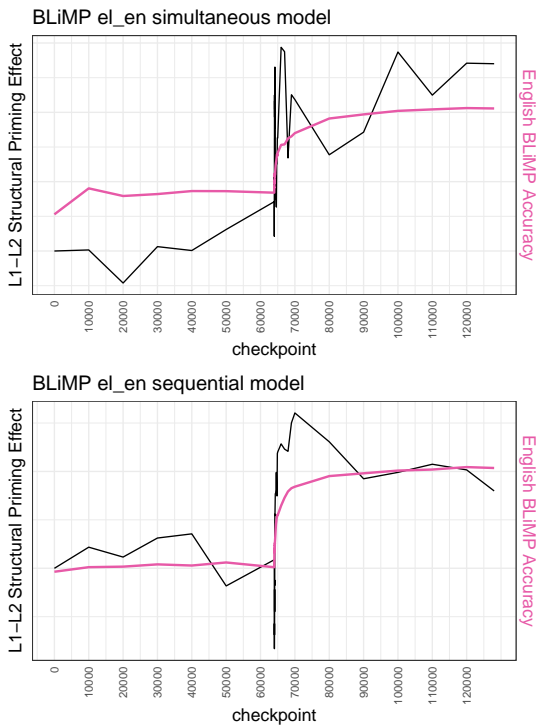

Figure E.12: Structural priming effect (black), plotted as the difference between match and mismatch conditions, and English BLiMP accuracy (pink) over the course of model training. Y-axes have been re-scaled for easier comparison.
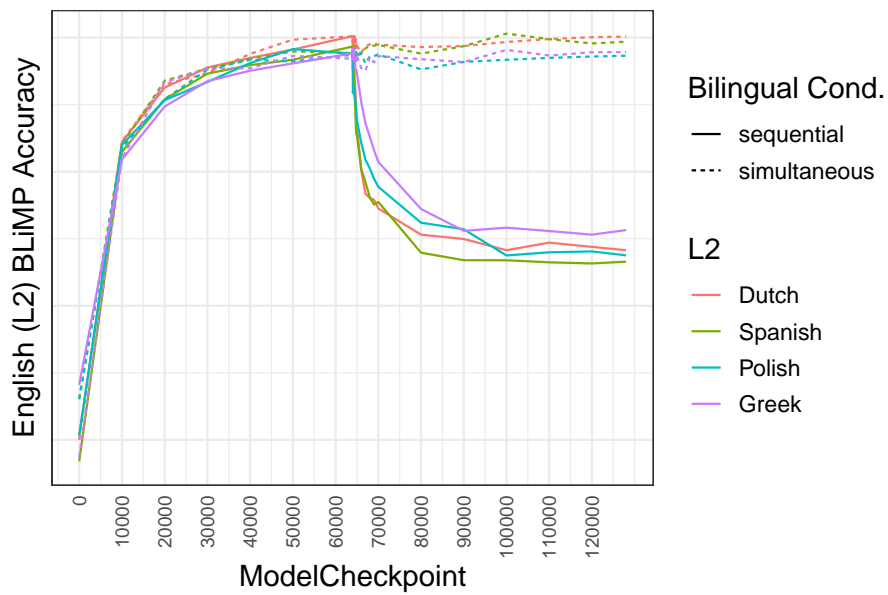
Figure F.13: English L1 models in both the sequential (solid lines) and simultaneous (dotted lines) conditions. BLiMP accuracy is plotted over the course of training.