Evaluating Typological Effects of L2 English in Language Model Benchmarks

Ellie Polyakova
University of Cambridge
August 2025

Abstract

We investigate second language acquisition (SLA) through the lens of language models by introducing and evaluating two datasets, BLiMP-fr and BLiMP-ru, to probe how interference and transfer occur when English is learnt as a second language. The datasets target eight linguistic phenomena, focusing on syntactic and morphological paradigms chosen for their typological traits to highlight potential cross-linguistic effects. We further introduce BabyLMs, pre-trained to emulate both monolingual and bilingual learners, to examine how L2 learning may differ across learner types. Our experiments reveal an asymmetry influenced by the typological relationship between L1 and L2: shared syntactic structures facilitate transfer, with English supporting French, whereas performance declines for Russian-specific phenomena.

1 Introduction

As first illustrated by BLiMP (Warstadt et al., 2023), minimal pairs are widely used to evaluate whether language models capture fine-grained grammatical contrasts. This entails a set of two grammatically similar sentences that differ by a single morphosyntactic feature affecting grammatical well-formedness, such as:

(1) a. *The cats chase the dog.*
b. **The cats chases the dog.*

Our own earlier experiments with MultiBLiMP (Jumelet et al., 2025) and XCOMPs (He et al., 2025) extended these insights to bilingual settings, where we observed well-documented challenges such as catastrophic forgetting in sequential bilingual models, as illustrated in Figure 1, and semantic interference across languages. The findings underscore the potential of extending minimal pairs for cross-lingual modelling. Motivated by prior research, which predominantly examined constraints within a single language, we introduce BLiMP-fr[1] and BLiMP-ru[2], two new minimal pair benchmarks for French and Russian. These resources extend the BLiMP paradigm to languages beyond English, enabling investigation of how models acquire and transfer knowledge across typologically distinct languages. In designing these datasets, we target linguistic phenomena most likely to reveal cross-lingual transfer processes in second language acquisition.

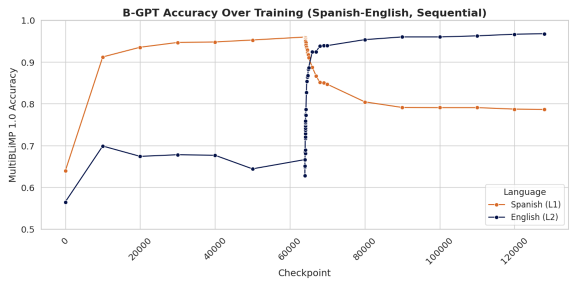

Figure 1: Bilingual-GPT (Arnett et al., 2025) Sequential Models evaluation on MultiBLiMP (Jumelet et al., 2025) with L1 Spanish and L2 English. The figure shows that sequential training leads to catastrophic forgetting, with accuracy on L1 decreasing as L2 is introduced.

To probe these effects, we train small-scale BabyLMs under both monolingual and bilingual conditions. This experimental setup allows us to directly test hypotheses regarding transfer and interference in multilingual learning. Furthermore,

[1] code: https://github.com/elliepreed/BLiMP-fr data: https://huggingface.co/datasets/elliepreed/BLiMP-fr
[2] code: https://github.com/elliepreed/BLiMP-ru data: https://huggingface.co/datasets/elliepreed/BLiMP-ru

we propose that these models can serve as student models of L2 learners, namely L2 English learners, providing a computational framework for examining how L1 properties influence L2 acquisition in both monolingual and bilingual learning contexts.

2 Minimal pair benchmark

Evaluating language models on a minimal pair basis has been a longstanding practice, targeting various linguistic phenomena. This method was first established by Warstadt et al. (2023) with the creation of BLiMP, an English benchmark comprising 67,000 sentence pairs across 67 syntactic paradigms. Since then, various monolingual models have been developed, such as TurBLiMP (Başar et al., 2025).

The evaluation method has expanded to a multilingual approach, as demonstrated by MultiBLiMP, which provides a dataset that covers two types of subject-verb agreement across 101 languages, with over 128,000 minimal pairs (Jumelet et al., 2025).

The primary focus of the previously discussed benchmarks was to create minimal pairs to evaluate the accuracy of language models in assigning a higher probability to grammatically correct sentences. This approach helps identify the linguistic phenomena that LLMs struggle with most. While inspired by benchmarks such as MultiBLiMP and TurBLimp, the phenomena in BLiMP-fr and BLiMP-ru are specifically adapted to the challenges of SLA. These benchmarks target French and Russian minimal pairs, with a focus on how interference and transfer arise when English is learned as a second language.

3 Typological differences

To design our experiments, we selected two typologically diverse L1s to gain deeper insights into L2 acquisition across distinct language families. French and Russian were chosen not only because they represent different families - Romance and Slavic, respectively - but also because the author's familiarity with both languages enables a more nuanced analysis of cross-linguistic transfer.

3.1 French

French is a Romance language with a relatively analytic morphosyntactic profile, though it retains fusional aspects. It follows a subject-verb-object (SVO) word order and exhibits a rich verbal inflectional system that marks tense, aspect, mood and agreement. Noun phrases are marked for gender and number, with agreement patterns extending to determiners and adjectives, providing clear syntactic cues (Fagyal et al., 2006).

(2) *Le* *facteur* *a*
DEF.M.SG mailman AUX.PRES.3SG
donné *la* *lettre* *à* *Maman.*
give.PART.PAST DEF.F.SG letter DAT Mom

'The mailman gave the letter to Mother.'

French is typologically closer to English, as both belong to the Indo-European language family and share similar morphosyntactic categories. This typological proximity is advantageous for BLiMP-fr research when evaluating LLMs. By comparing model performance on French-derived datasets with that on a more typologically distant L1 such as Russian, we can better isolate the impact of cross-linguistic interference and transfer.

3.2 Russian

Russian presents a particularly compelling case for BLiMP-style evaluation due to its flexible word order and rich morphological system. Russian is a Slavic language with a highly synthetic and fusional morphosyntactic profile (Reynolds, 2016). It features relatively free word order, extensive case marking across six grammatical cases and a complex verbal morphology (Dyakonova, 2009). For example:

(3) *Ol'ga* *svarila* *pel'meni.*
Olga.NOM cook.PST.FEM pelmeni.ACC

'Olga cooked pelmeni (Russian dumplings).'

These features create a high degree of morphological and syntactic variability compared to English. Due to its typological distinctiveness, Russian serves as an ideal L1 for evaluating cross-linguistic transfer and assessing the effects of interference. The contrast between a typologically similar language (French) and a typologically distant one (Russian) enables BLiMP-style benchmarks to reveal how L1 properties shape syntactic and morphological learning in language models.

4 BLiMP-fr

The development of the BLiMP-fr benchmark and the selection of specific linguistic phenomena were

Phenomenon	Minimal pair	Translation
Adjective-noun agreement	*Les [anciens/*anciennes] oiseaux sont sur la table.*	The old birds are on the table.
Anaphor agreement	*Je* [*me* / **se*] lave.	I wash [myself/*herself].
Clitic placement	*Je le vois / *Je vois le.*	I see it/*I it see.
Determiner-noun agreement	*La robe est dans* [*le* / **la*] placard.	The dress is in the closet.
Negation	*Je ne mange pas/*Je mange pas.*	I don't eat.
Past participle agreement	*Elle est* [*arrivée/*arrivé*] ce matin.	She arrived this morning.
Subject-verb agreement	*John* [*aime* / **aiment*] Marie.	John [loves/*love] Marie.
Subjunctive mood	*Pourvu qu'il* [*ait/*a*] raison.	As long as he is right.

Table 1: Glossed minimal pairs for each phenomenon in BLiMP-fr. For the purpose of this benchmark, grammatical sentences follow pedagogical rules rather than colloquial speech. Most of these phenomena have no direct ungrammatical counterparts in English, as these structures are absent from the language.

motivated by the need for a metric that offers insights into SLA. Several of the targeted phenomena reflect syntactic structures present in French but absent in English, thereby providing a means to evaluate cross-linguistic transfer in L2 learning. The following section presents a brief linguistic overview of the minimal pairs used in the benchmark.

To justify our choice of minimal pairs, we draw upon descriptions of French syntax from learner-oriented pedagogical materials, which highlight constructions that differ systematically from English. While these sources sometimes simplify aspects of French, they offer consistent structural explanations that motivate our selection of the phenomena described below.

4.1 Phenomena

This dataset comprises eight linguistic phenomena, each represented by 1,000 minimal pairs.

ANAPHOR AGREEMENT In French, reflexive pronouns must agree in person and number with the subject (Mazet, 2013). Ungrammatical examples in the dataset feature mismatches between the reflexive and its antecedent, violating agreement constraints.

ADJECTIVE-NOUN AGREEMENT In French, adjectives must agree with the noun they modify in both gender and number, unlike in English (Mazet, 2013).Ungrammatical examples in the dataset involve mismatches in agreement between the adjective and the noun.

CLITIC PLACEMENT In French, object clitic pronouns must precede the verb, whereas in English, object pronouns typically follow it. Ungrammatical examples in the dataset involve incorrect placement of the clitic pronoun.

DETERMINER-NOUN AGREEMENT In French, determiners must agree with the noun in gender and number, whereas in English, they are invariant (Mazet, 2013). Ungrammatical examples involve mismatches in agreement between the determiner and the noun.

NEGATION In French, negation is typically expressed using two particles: 'ne' and 'pas'. In contrast, English expresses negation with a single marker, 'not' (Mazet, 2013). Ungrammatical examples omit the 'ne' particle, violating the bipartite negation structure in standard French.

PAST PARTICIPLE AGREEMENT In French, the auxiliary verb 'être' is used to form the passé composé with certain verbs and triggers agreement between the past participle and the subject - a phenomenon not present in English (Mazet, 2013). Ungrammatical examples involve incorrect past participle agreement with the subject.

SUBJECT-VERB AGREEMENT In French, verbs agree with the subject across all grammatical persons, whereas in English, overt agreement is limited to the third-person singular form. Ungrammatical examples involve mismatches in subject-verb agreement, typically in number or person.

SUBJUNCTIVE MOOD The subjunctive is commonly used in French to express uncertainty, wishes and other subjunctive contexts, whereas in English, the subjunctive is rarely employed. (Mazet, 2013). Ungrammatical examples omit the subjunctive where it is grammatically required.

These typological differences make the selected phenomena especially useful for evaluating L2 acquisition, as they help reveal whether a model adheres to French grammatical rules or whether performance is impeded by interference from English.

Features that are absent in English, such as noun gender and past participle agreement, serve as valuable evaluation points, enabling us to assess how well models adapt to cross-linguistic variation and overcome the morphological and syntactic differences between the languages.

While some structures overlap across both languages, this contrastive setup provides a necessary baseline for evaluating the extent of transfer and interference. Overall, these phenomena present a compelling challenge for SLA benchmarking, offering insight into how typological distance shapes syntactic and morphological generalisation in language models.

5 BLiMP-ru

Building on BLiMP-fr, this evaluation benchmark extends the work of Taktasheva et al. (2024) on ru-BLiMP and and is designed to evaluate cross-linguistic transfer. The following phenomena have been specifically selected to capture challenges relevant to SLA. Following the rationale used for BLiMP-fr, we draw on pedagogical materials that offer clear, learner-oriented descriptions of Russian syntax to guide the construction of minimal pairs.

5.1 Phenomena

This dataset contains eight phenomena, each represented by 1,000 minimal pairs.

ACCUSATIVE MARKING Russian has accusative case marking, primarily to indicate direct objects, whereas in English this is obsolete (Kaufman et al., 2006). Ungrammatical sentences apply incorrect case forms, such as using the nominative instead of the accusative.

ASPECT Russian distinguishes between imperfective (ongoing) and perfective (completed) aspectual forms (Kaufman et al., 2006). Ungrammatical examples involve mismatches between aspect and temporal adverbs (e.g. using the perfective with ongoing contexts).

COPULAR VERB OMISSION In Russian, the present-tense form of the verb 'to be' is not expressed, whereas in English the copula must be overtly realised (Kaufman et al., 2006). Ungrammatical examples retain the copula where it should be omitted.

GENITIVE NEGATION Under negation, the direct object of a verb in Russian is often expressed in the genitive case (Kaufman et al., 2006). Ungrammatical examples involve incorrect case marking.

NOMINAL DERIVATION Russian nouns can be transformed into adjectival forms to modify other nouns, a structure that contrasts with English, where bare attributive nouns are more commonly used (Corbett, 2004). Ungrammatical examples involve the use of a bare noun where a derived adjective is required.

NUMBER AGREEMENT In Russian, nouns take different morphological endings depending on the case and must agree in number with quantifiers (Kaufman et al., 2006). Ungrammatical examples involve the use of singular forms where plural agreement is required.

THIRD PERSON AGREEMENT In both Russian and English, finite verbs must agree with their subjects in both person and number. Ungrammatical examples feature incorrect subject-verb agreement.

TRANSITIVITY ERROR Both English and Russian distinguish between transitive and intransitive verbs. Ungrammatical examples pair intransitive verbs with direct objects, violating subcategorisation constraints.

The selected phenomena balance typological similarity and divergence between Russian and English. This design enables us to evaluate how structural differences influence minimal pair evaluation when English is acquired as a second language.

6 Benchmark creation

In the creation of BLiMP-fr and BLiMP-ru, ten templates were manually constructed for each linguistic phenomenon to provide clear guidelines. This initial step ensured that each minimal pair differed only in the targeted grammatical feature.

These manually designed templates were then expanded to 1,000 minimal pairs per phenomenon

Phenomenon	Minimal pair	Translation
Accusative marking	Я пью [молоко/*молоку] из стакана.	I am drinking milk from the glass.
Aspect	Она два часа [читала/*прочитала] книгу.	She [was reading/*read] the book for two hours.
Copular verb omission	Она [*есть] дома.	She [is/*-] at home.
Genitive negation	Я [новостей/*новости] не смотрю.	I don't watch the news.
Nominal derivation	Я люблю [творожную/*творог] запеканку.	I like [curd-ADJ/*curd-N] zapekanka (Russian cottage cheese bake).
Number agreement	У меня есть три [книги/*книга].	I have three [books/*book].
Third-person agreement	Он [хочет/*хочу] читать книгу.	He [wants/*want] to read the book.
Transitivity error	Мальчик спит [*подушку].	The boy sleeps [*the pillow].

Table 2: Glossed minimal pairs for each phenomenon in BLiMP-ru. For the purpose of this benchmark, grammatical sentences follow pedagogical rules rather than colloquial speech. Phenomena such as accusative marking and genitive negation have no direct ungrammatical counterparts in English, as these structures are absent from the language.

through automatic augmentation. A Python script was developed to systematically apply each template, iterating over lexical items and syntactic configurations to generate grammatically controlled sentence pairs. This process formed the foundation of the final datasets used for evaluation in BLiMP-fr and BLiMP-ru.

7 Experimental setup

We evaluate models on BLiMP-fr and BLiMP-ru to investigate the influence of English as a second language in multilingual contexts. To model different second-language acquisition scenarios, we propose an experimental setup that simulates both adult L2 acquisition and bilingual exposure. Specifically, we compare:

- Monolingual models fine-tuned on English, simulating adult L2 learners, and
- Bilingual models trained on both L1 and L2, simulating simultaneous and sequential bilinguals.

This design allows us to evaluate how typological differences between L1 and English influence syntactic generalisation, particularly under varying levels of English exposure, reflecting real-world SLA contexts.

7.1 Monolingual BabyLMs

We develop monolingual BabyLMs for French[3] and Russian[4], using the training code[5] provided. These models are pretrained solely on L1 data and serve as baselines for performance on the BLiMP-fr and BLiMP-ru benchmarks. This allow us to evaluate whether L1-specific syntactic properties can be successfully modelled in isolation, prior to any English exposure.

7.2 Bilingual BabyLMs

Adult L2 learner To simulate adult second language learners[6], the monolingual BabyLMs were subsequently fine-tuned on English. In this setup, the pretraining phase reflects L1 acquisition during the critical period, while fine-tuning on English represents the delayed onset of second language acquisition (Slabakova, 2013).

Bilingual learner To simulate bilingual acquisition, we draw inspiration from the B-GPT framework proposed by Arnett et al. (2025), implementing two bilingual acquisition scenarios for each L1,

[3]https://huggingface.co/elliepreed/french-babylm-urop-Ellie

[4]https://huggingface.co/climb-mao/russian-babylm-urop-shivan

[5]code: https://github.com/suchirsalhan/babylm-tutorial

[6]FR: https://huggingface.co/elliepreed/french-babylm-ft RU:https://huggingface.co/elliepreed/russian-babylm-ft

French[7] and Russian[8], with English as the shared L2. Within each L1 group, the key distinction lies in the type of bilingualism:

- Simultaneous bilingual condition: Models are exclusively trained on L1 during the first half of training, followed by a balanced mix of L1 and L2 in the second half.
- Sequential bilingual condition: Models are trained only on the L1 during the first half, and only on the L2 during the second half.

Training data for these bilingual models was derived by merging resources from Geertzen et al. (2014) and Nicholls et al. (2024).

This design enables us to examine both the influence of bilingual type (simultaneous vs. sequential) and the role of English as the second language on syntactic transfer and interference.

8 Results

8.1 BLiMP-fr evaluation

Phenomenon	FR Babylm	FR→EN ft.	Bil. Sim.	Bil. Seq.
Adj. agr.	0.901	0.615	0.900	0.919
Anaphor agr.	0.857	0.723	0.910	0.908
Clitic place.	0.874	0.681	0.938	0.873
Determiners	0.979	0.689	1.000	0.999
Negation	0.707	0.076	0.917	0.844
Past participle agr.	0.829	0.661	0.828	0.851
Subj.verb agr.	0.690	0.745	0.964	0.989
Subjunctive	0.629	0.111	0.750	0.755
Macro avg.	0.808	0.538	0.901	0.892

Table 3: BLiMP-fr accuracies evaluated on French BabyLM, fine-tuned BabyLM on English, simultaneous and sequential BabyLM models; L1 French L2 English (green = transfer, red = interference, yellow = no change).

Model evaluation The performances of the models across the selected linguistic phenomena are summarised in Table 3. The results reveal that bilingual models outperform both the monolingual French BabyLM and the fine-tuned L2 model. In terms of overall average accuracy, the simultaneous bilingual model performs best, with improved accuracies across nearly all phenomena. The sequential bilingual model performs comparably, reaching a similar overall accuracy and closely matching the improvements of the simultaneous model. In contrast, the fine-tuned model shows a marked decline in performance, with significantly lower accuracy across most categories.

Phenomena evaluation At the level of individual phenomena, subject-verb agreement is the only case where all models, including the fine-tuned and bilingual variants, outperform the French BabyLM baseline. For most other phenomena, consistent gains are observed only in the bilingual conditions. In particular, anaphor agreement, negation and the subjunctive mood show substantial improvements under bilingual training, mirroring the strong upward trend seen in subject-verb agreement.

Phenomena involving determiner-noun and adjective-noun agreement remain consistently high across all bilingual conditions, with little variation between training types. Past participle agreement follows a similar pattern, though the improvements are less pronounced.

A notable divergence is observed in clitic placement: while the simultaneous bilingual model shows a significant gain, the sequential bilingual model maintains performance close to the monolingual baseline.

8.2 BLiMP-ru evaluation

Phenomenon	RU Babylm	RU→EN ft.	Bil. Sim.	Bil. Seq.
3rd inflect.	0.689	0.617	0.773	0.667
Acc. marking	0.817	0.633	0.610	0.635
Aspect	0.866	0.844	0.617	0.743
Copular omission	1.000	1.000	0.966	0.965
Genitive	0.773	0.844	0.654	0.661
Intransitive	1.000	1.000	1.000	1.000
Nominal deriv.	0.0901	0.0911	0.2931	0.5614
Number agr.	0.712	0.544	0.542	0.580
Macro avg.	0.743	0.697	0.682	0.727

Table 4: BLiMP-ru accuracies evaluated on Russian BabyLM, fine-tune BabyLM on English, simultaneous and sequential BabyLM models; L1 Russian L2 English (green = transfer, red = interference, yellow = no change).

Model evaluation The performances of the models across the selected linguistic phenomena are summarised in Table 4. In contrast to BLiMP-fr, where bilingual training consistently improved over the baseline, the results for BLiMP-ru show no clear benefit from either bilingual training or fine-tuning. All models underperform relative to

[7]simultaneous: https://huggingface.co/elliepreed/bgpt-french-english sequential: https://huggingface.co/elliepreed/french_english_sequential

[8]simultaneous: https://huggingface.co/elliepreed/russian_english_simultaneous sequential: https://huggingface.co/elliepreed/russian_english_sequential

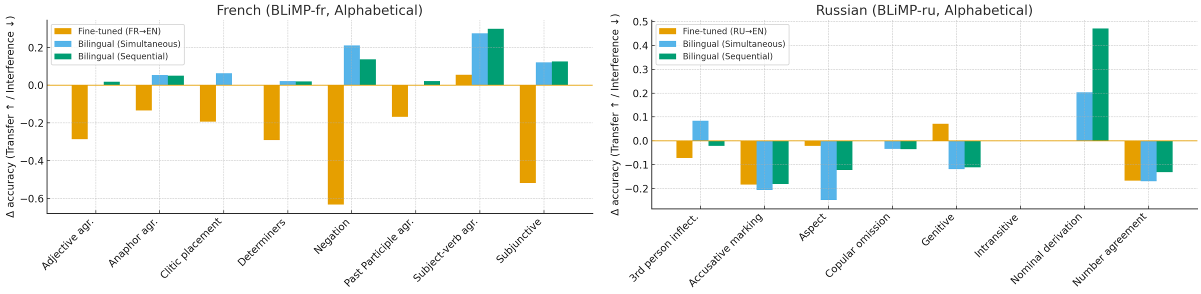

Figure 2: Per-phenomenon effects of English exposure, illustrating cases of cross-lingual transfer (positive change) and interference (negative change) relative to the BabyLM baseline.

the Russian BabyLM baseline.

Among the evaluated models, the sequential bilingual model achieves the highest overall accuracy, followed closely by the fine-tuned model and then the simultaneous bilingual model. Notably, the fine-tuned model slightly outperforms the simultaneous bilingual model - a reversal of the pattern observed in BLiMP-fr, where fine-tuning resulted in the lowest performance.

Phenomena evaluation Performance across individual phenomena reveals a general downward trend, though with important variation. Accusative marking, aspect and number agreement all show notable declines across all models. In contrast, transitivity and nominal derivation remain relatively stable, with bilingual training producing slight improvements in these areas.

For high-accuracy phenomena such as copular omission, performance is largely preserved across models. Interestingly, the fine-tuned model achieves the strongest results on this phenomenon, while both bilingual models show small performance drops. Third-person inflection and genitive under negation also exhibit only minor decreases, with the simultaneous bilingual and fine-tuned models, respectively, outperforming the baseline in each case.

9 Discussion

9.1 Cross-model evaluation

The results align with expectations from research in human bilingualism. In simultaneous bilingual learning, where both languages are acquired in parallel, balanced exposure typically leads to robust linguistic competence in both languages (Rivero, 2018). This pattern is mirrored in our findings: in BLiMP-fr, both bilingual models outperform the monolingual French BabyLM, and in BLiMP-ru, bilingual models perform comparably to the baseline. This suggests that bilingual training enables the model to accommodate both languages while largely preserving L1-specific competence.

By contrast, the fine-tuned models, designed to approximate adult L2 learners, show more variable performance. In human SLA, adult learners often struggle with L1 transfer, finding it difficult to overcome this linguistic barrier and adapt to new morphologic and syntactic dependencies (Wu et al., 2021). The decline in the fine-tuned models' accuracy is consistent with catastrophic forgetting, a phenomenon previously observed in sequential training setups by Arnett et al. (2025). The findings suggest that post-hoc English exposure overwrites much of the models' L1 knowledge. Since fine-tuning involved exclusive exposure to English, the models likely developed more refined L2 representations at the expense of L1 knowledge. Unlike human adult learners, who typically retain entrenched L1 representations throughout life, these models appear more vulnerable to complete L1 overwriting, especially when there is no L1 exposure during fine-tuning. This presents a more challenging environment for constructing computational representations of adult L2 learners. Unlike human learners, who - even under conditions of extensive immersion - very rarely achieve native-like proficiency due to the persistence of entrenched L1 representations, neural models are often capable of entirely overwriting their L1 when exposed exclusively to L2 input (Wu et al., 2021). This effect more closely resembles language dominance shifts observed in early bilinguals, rather than the persistent L1 interference seen in mature L2 learners (Hammer, 2021).

Interestingly, our results diverge from previous findings in one key area. While Arnett et al. (2025) report severe forgetting in sequential bilingual models, our sequential bilinguals - particularly in Russian - did not suffer significant performance degradation. In fact, the Russian sequential model outperformed the simultaneous one. One possible explanation is the greater typological distance between Russian and English, which may protect against overwriting due to reduced grammatical overlap. Conversely, French and English share many syntactic structures, particularly subject-verb agreement, increasing the potential for interference and loss. Furthermore, the Russian fine-tuned model did not degrade as sharply as the French one. This again suggests that typological proximity may heighten vulnerability to forgetting in sequential and fine-tuned settings. However, it is also important to note that the linguistic phenomena targeted in BLiMP-fr were selected to highlight areas of typological overlap, and therefore, these results may not generalise to all aspects of French grammar.

9.2 Phenomenon-level effects

Our results reveal a striking asymmetry between French and Russian: while bilingual training consistently improves performance over the BabyLM baseline in French, all Russian models exhibit a decline in accuracy. This contrast can be attributed to typological distance: English and French share core morphosyntactic features, whereas Russian diverges significantly in both morphology and syntax (Lee, 2022). As a result, bilingual training facilitates positive transfer in French, but leads to increased interference in Russian.

A more detailed analysis (see Figure 2) indicates that phenomena shared across languages in structurally similar ways are indeed beneficial for model evaluation. The strongest evidence of transfer appears in French, particularly with subject-verb agreement, negation and the subjunctive, where the bilingual models significantly outperform the base model. These phenomena, while not identical, overlap structurally with English causing a positive transfer effect. For example, subject-verb agreement is productively used in both languages, which may account for the substantial improvements in performance (Franck et al., 2002). The subjunctive, though less frequent in English, shares features with English modal constructions and clausal complementation, offering partial structural cues (Quer, 2009). Negation similarly exhibits evidence of transfer. Although surface constructions differ slightly - such as 'ne. . . pas' in French - dialectal varieties of English, including negative concord in African American Vernacular English, offer comparable structures, potentially reducing cross-linguistic interference (Labov, 1972). Other high-performing French phenomena include determiner-noun agreement and anaphor agreement, both of which are productively realised in both French and English.

In contrast, the Russian results underscore the limits of cross-linguistic transfer. Phenomena such as intransitive verb usage, nominal derivation and third-person inflection remain relatively stable, likely due to partial structural overlap or low morphological complexity. However, most other phenomena reveal significant performance degradation under bilingual and fine-tuned conditions. In particular, accusative marking, aspect and number agreement show the most pronounced declines. These categories rely heavily on rich case morphology and verbal inflection, features that are largely absent in English (Blake, 2001). The lack of morphosyntactic marking in English for these features likely explains the heightened interference observed.

These findings confirm that transfer effects are highly dependent on typological similarity. Positive transfer is most likely when the L1 and L2 share grammatical categories and surface forms; conversely, typologically distant features - especially those grounded in morphological complexity - are more vulnerable to interference and model degradation.

9.3 Transfer in semantic evaluation (XCOMPs)

To further investigate the impact of typological similarity on cross-linguistic transfer, we evaluated our BabyLMs using XCOMPs (He et al., 2025), a multilingual benchmark of conceptual minimal pairs. This allowed us to assess whether transfer effects extend beyond syntax into the semantic domain. The results are summarised in Table 5.

In line with our syntactic findings from BLiMP-fr, the French bilingual models consistently outperform the monolingual BabyLM, while the fine-tuned model shows a minor performance decline. However, the Russian results diverge from the BLiMP-ru pattern: both bilingual models achieve modest improvements over the monolingual baseline, and fine-tuning has negligible impact, with

Model	French	Russian
BabyLM	52.05 (7479/14368)	48.66 (6992/14368)
Fine-tuned EN	50.35 (7235/14368)	48.55 (6976/14368)
Bilingual Sim.	54.64 (7850/14368)	50.79 (7297/14368)
Bilingual Seq.	54.64 (7850/14368)	50.62 (7273/14368)

Table 5: XCOMPs accuracies (%). Colors indicate change relative to BabyLM: green = transfer (improvement), red = interference (decline).

only a slight decrease in accuracy.

These findings suggest that semantic transfer is more resilient to typological distance than syntactic transfer, with bilingual training providing benefits even in Russian, where syntactic interference is most pronounced. Unlike morphosyntactic features, which are highly sensitive to structural divergence, conceptual representations may generalise more easily across typologically distinct languages.

10 Conclusion

This paper introduced BLiMP-fr and BLiMP-ru, two syntactic benchmarks designed to evaluate English L2 acquisition in bilingual language models. Together, they offer a comprehensive framework for exploring how typological similarity shapes cross-linguistic transfer. Our findings reveal that typologically aligned languages, such as French and English, facilitate positive transfer, with bilingual French models demonstrating substantial improvements over monolingual baselines. In contrast, Russian models, typologically distant from English, exhibited greater interference.

Fine-tuned models, intended to simulate adult L2 learners, consistently suffered from catastrophic forgetting, effectively overwriting previously acquired L1 knowledge. These results suggest that simultaneous bilingual training better preserves grammatical competence in both languages. Supplementary evaluation using XCOMPs further demonstrated that transfer is most robust in syntactic benchmarks, while semantic transfer appeared more resilient across typological boundaries but less pronounced overall.

For future work, it will be important to test whether these trends hold in larger pretrained models, extending the analysis beyond the BabyLM architecture. Further investigation of fine-tuning strategies may help mitigate catastrophic forgetting and clarify whether positive transfer requires simultaneous training on both languages. Additionally, incorporating varying levels of L1/L2 dominance could provide a more nuanced account of bilingualism in computational models, bringing them closer to the dynamics observed in human language learners.

Limitations

While BLiMP-fr and BLiMP-ru offer a novel and comprehensive framework for evaluating cross-linguistic transfer and interference in bilingual language models, several limitations should be acknowledged.

First, in constructing the BLiMP-fr and BLiMP-ru datasets, we implemented a custom augmentation pipeline in Python to expand the number of minimal pairs from 10 templates to 10,000 instances. Due to constraints in time and experience, the augmentation process introduced some inconsistencies, including occasional syntactic mismatches and semantic incoherence. Although a subset of the data was manually reviewed, a full quality assurance process was not feasible within the scope of this project. Nonetheless, the dataset remains a valuable resource for benchmarking purposes. Future work should prioritise a more rigorous validation phase, including refinement of the augmentation techniques and human evaluation, to ensure higher levels of consistency and accuracy.

Second, the template design was based on the author's non-native knowledge of French and Russian. This may have resulted in subtle grammatical inaccuracies in some minimal pair items. While the use of pedagogical grammar sources helped to guide design decisions, native speaker validation would enhance linguistic reliability.

Third, we initially considered including binding as one of the target phenomena, but ultimately excluded it due to the interpretive complexity involved in evaluating ungrammatical constructions. Binding constraints are highly context-dependent, and their violation often relies on nuanced pragmatic judgements. Given that LLMs struggle to reliably capture such context-sensitive dependencies, we deemed this phenomenon unsuitable for our evaluation.

Finally, the Russian dataset used for training was substantially smaller than the French dataset. This imbalance may have influenced the performance outcomes and makes direct comparisons between the two languages less robust. For future work, ensuring comparable data sizes across language pairs would provide more accurate results.

Despite these limitations, this study presents an initial framework for evaluating L2 English acquisition in bilingual language models through minimal pair analysis, offering insights into the representation and transfer of grammatical knowledge across typologically diverse languages.

Acknowledgments

I would like to thank Suchir Salhan and Andrew Caines for their invaluable guidance, advice and support throughout this project. I am especially grateful to Suchir Salhan for providing the code used to train the models. I also extend thanks to Shivan Arora for his contributions to BLiMP-ru and for his role in advancing the development of the Russian BabyLM. This research was supported by Cambridge University Press & Assessment, and was made possible by a donation of a Titan X Pascal GPU from NVIDIA Corporation.

References

Catherine Arnett, Tyler A. Chang, James A. Michaelov, and Benjamin K. Bergen. 2025. On the Acquisition of Shared Grammatical Representations in Bilingual Language Models.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs.

Barry J. Blake. 2001. *Case*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2 edition.

Greville G. Corbett. 2004. The Russian Adjective: A Pervasive yet Elusive Category. In R. M. W. Dixon and Alexandra Y. Aikhenvald, editors, *Adjective Classes: A Cross-Linguistic Typology*, pages 201–226. Oxford University Press, Oxford.

Marina Dyakonova. 2009. *A Phase-Based Approach to Russian Free Word Order*. Ph.D. thesis, University of Amsterdam, Utrecht.

Zsuzsanna Fagyal, Douglas Kibbee, and Frederic Jenkins. 2006. *French: A Linguistic Introduction*. Cambridge University Press.

Julie Franck, Gabriella Vigliocco, and Janet Nicol. 2002. Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17(4):371–404.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2014. Automatic Linguistic Annotation of Large-Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. Millar, K. I. Martin, C. M. Eddington, A. Henery, N. M. Miguel, and A. Tseng, editors, *Selected Proceedings of the 2012 Second Language Research Forum*, pages 240–254, Somerville, MA. Cascadilla Proceedings Project.

Kate Hammer. 2021. Hammer, K. (2021). Shift in language dominance in bilinguals: An acculturation perspective. *Sociolinguistic Studies*, 15(2-4):299–322, 07.

Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Adrian Florea, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, Hinrich Schütze, and Nima Mesgarani. 2025. XCOMPS: A Multilingual Benchmark of Conceptual Minimal Pairs.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs.

Andrew Kaufman, Serafima Gettys, and Nina Mayo. 2006. *Russian For Dummies*. Wiley, Hoboken, NJ.

William Labov. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press, Philadelphia, PA.

Yeo Y. Lee. 2022. A conceptual analysis of typological distance and its potential consequences on the bilingual brain. *International Journal of Bilingual Education and Bilingualism*, 25(9):3333–3346.

Véronique Mazet. 2013. *French Grammar for Dummies*. For Dummies. John Wiley & Sons, Indianapolis, IN. Includes index.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. *The Write amp; Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English*.

Josep Quer. 2009. Twists of mood: The distribution and interpretation of indicative and subjunctive. *Lingua*, 119:1779–1787, 12.

Robert J. Reynolds. 2016. *Russian Natural Language Processing for Computer-Assisted Language Learning: Capturing the Benefits of Deep Morphological Analysis in Real-Life Applications*. Ph.D. thesis, UiT: Arctic University of Norway, Tromsø, Norway, February. Printed by Tromsprodukt AS.

Rowena Rivero. 2018. Simultaneous Bilingual Child: A Language Acquisition Study. *Scientia - The International Journal on the Liberal Arts*, 7, 03.

Roumyana Slabakova. 2013. Adult second language acquisition: A selective overview with a focus on the learner linguistic system. *Linguistic Approaches to Bilingualism*, 3(1):48–72.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2023. BLiMP: The Benchmark of Linguistic Minimal Pairs for English.

Fuyun Wu, Jun Lyu, and Yanan Sheng. 2021. Effects of L1 Transfer Are Profound, Yet Native-Like Processing Strategy Is Attainable: Evidence From Advanced Learners' Production of Complex L2 Chinese Structures. *Frontiers in Psychology*, 12:794500.