

Evaluating the Cross-Lingual Syntactic Capabilities of Language Models

Suchir Salhan

sas245@cam.ac.uk.¹

Computer Laboratory, University of Cambridge, UK

L95 2025/26

27 October 2025

¹ *An early draft of these slides was prepared Dr David Strohmaier based on a handout by Suchir Salhan*

Grammaticality, Multilinguality and LLMs

Minimal Pair Evaluation

Syntactic Theory and BLiMP

Minimal Pair Datasets Beyond English

Limitations of Minimal Pairs

Linguistic Interpretability

Minimal Pair Evaluation

Minimal pairs datasets consists of contrasting grammatical and ungrammatical sentences.

Description

- ▶ Should diverge minimally.
- ▶ Sentences should be of same length.

Calculating Scores for Minimal Pairs

- ▶ Autoregressive models: apply chain rule
- ▶ Masked language models: Sentence pseudo-log-likelihood

$$\text{PLL}(\mathbf{W}) := \sum_{t=1}^{|\mathbf{W}|} \log P_{\text{MLM}}(\mathbf{w}_t | \mathbf{W}_{\setminus t}; \Theta) \quad (1)$$

BLiMP Dataset

BLiMP by Warstadt et al. (2020)

Description

- ▶ Sentence pairs generated from abstract grammar.
- ▶ Sentences exemplify 12 principles of syntax and 67 syntactic paradigms.
- ▶ Human benchmark included.

BLiMP Dataset: The 12 Phenomena

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	<i>Many girls insulted <u>themselves</u>.</i>	<i>Many girls insulted <u>herself</u>.</i>
ARG. STRUCTURE	9	<i>Rose wasn't <u>disturbing</u> Mark.</i>	<i>Rose wasn't <u>boasting</u> Mark.</i>
BINDING	7	<i>Carlos said that Lori helped <u>him</u>.</i>	<i>Carlos said that Lori helped <u>himself</u>.</i>
CONTROL/RAISING	5	<i>There was <u>bound</u> to be a fish escaping.</i>	<i>There was <u>unable</u> to be a fish escaping.</i>
DET.-NOUN AGR.	8	<i>Rachelle had bought that <u>chair</u>.</i>	<i>Rachelle had bought that <u>chairs</u>.</i>
ELLIPSIS	2	<i>Anne's doctor cleans one <u>important</u> book and Stacey cleans a few.</i>	<i>Anne's doctor cleans one book and Stacey cleans a few <u>important</u>.</i>
FILLER-GAP	7	<i>Brett knew <u>what</u> many waiters find.</i>	<i>Brett knew <u>that</u> many waiters find.</i>
IRREGULAR FORMS	2	<i>Aaron <u>broke</u> the unicycle.</i>	<i>Aaron <u>broken</u> the unicycle.</i>
ISLAND EFFECTS	8	<i>Which <u>bikes</u> is John fixing?</i>	<i>Which is John fixing <u>bikes</u>?</i>
NPI LICENSING	7	<i>The truck has <u>clearly</u> tipped over.</i>	<i>The truck has <u>ever</u> tipped over.</i>
QUANTIFIERS	4	<i>No boy knew <u>fewer than</u> six guys.</i>	<i>No boy knew <u>at most</u> six guys.</i>
SUBJECT-VERB AGR.	6	<i>These casseroles <u>disgust</u> Kayla.</i>	<i>These casseroles <u>disgusts</u> Kayla.</i>

Table: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. *N* is the number of 1,000-example minimal pair paradigms within each broad category. Table and description copied from original paper by Warstadt et al. (2020).

Agreement

BLiMP contains different types of agreement:

- ▶ Subject-verb agreement
- ▶ Anaphoric agreement
- ▶ Determiner-noun agreement
- ▶ Also included in irregular forms

Filler-Gap Dependencies & Island Effects

Filler-Gap dependencies

Filler-Gap Dependencies arise from phrasal movement in, e.g., wh-questions.

- ▶ subject gaps
- ▶ object gaps

Island effects

- ▶ Adjunct islands
- ▶ Complex NP islands
- ▶ Coordination islands
- ▶ Left branch islands

Referential Expressions

- ▶ Anaphors, i.e. reflexive pronouns such as themselves
- ▶ Pronouns, e.g. they
- ▶ R-expressions, i.e. noun phrases that refer to an entity in the world

BLiMP's binding dataset targets properties of the structural relationship between a pronoun and its antecedent.

Control and Raising

- ▶ Control and raising refers to two types of constructions.
- ▶ They highlight syntactic-semantic differences between types of predicates.
- ▶ BLiMP covers only simple cases of subject raising and subject control.

Example

*William has **declared**/***obliged** there to be no guests getting fired.*

Ellipsis

- ▶ BLiMP constrained by equal sentence length requirement.
- ▶ Covers special cases of NP ellipsis.

Example

- ▶ *Brad passed one big museum and Eva passed several.*
- ▶ **Brad passed one museum and Eva passed several big.*

Syntax-Semantic Interface

- ▶ Argument structure
- ▶ NPI licensing
- ▶ Quantifiers

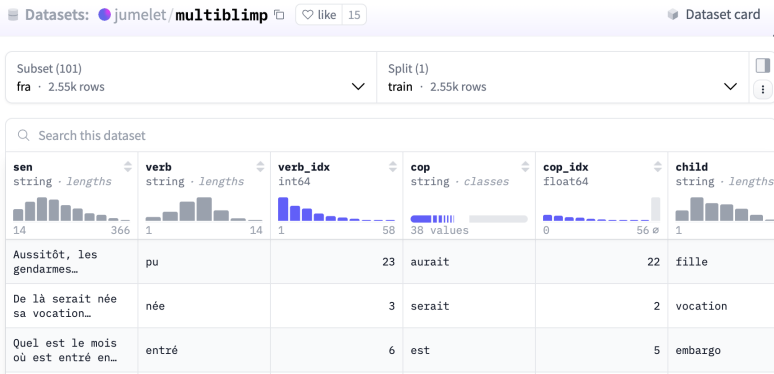
Minimal Pair Datasets Beyond English

Name	Languages
CLAMS	French and German
JBLiMP	Japanese
SLING	Chinese
TurBLiMP	Turkish
BLiMP-NL	Dutch
MultiBLiMP	101 languages

Table: Other minimal pair datasets.

MultiBLiMP 1.0

A modern, large-scale repurposing of Dependency Grammars for LLMs. More on this next week!



MultiBLiMP 1.0 on HuggingFace

MultiBLiMP 1.0: Data Construction

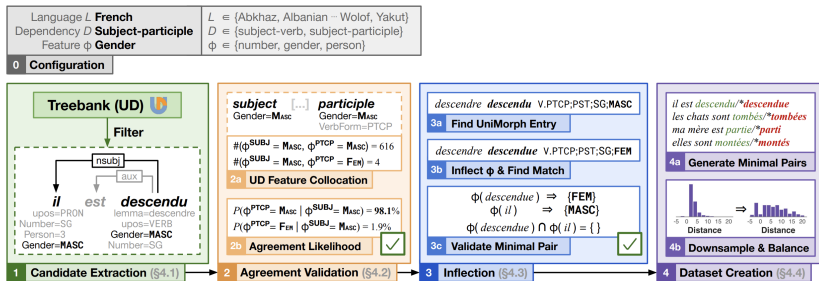


Figure 2: Pipeline of the minimal pair creation procedure of MultiBLiMP 1.0

Figure from Jumelet et al (2025) MultiBLiMP 1.0

MultiBLiMP: Example of the Challenges

*??Dadurch **sollten** Jahre später das Königreich Powys an die Herrscher des Königreiches Gwynedd fallen.*

This way, the Kingdom Powys should. PL years later fall to the rulers of the Kingdom Gwynedd.

Context

- ▶ Switching singular→plural for verbs
- ▶ Example uses Konjunktiv I (subjunctive mood), which is supposed to be used for indirect speech
- ▶ Ongoing language change
- ▶ There are multiple such cases in the dataset!

Beyond BLiMP-style Datasets

- ▶ Human abilities might generalise in ways not tested for by minimal pair datasets
- ▶ Issue 1: No strict grammaticality threshold (**gradient** of acceptability judgements)
- ▶ Issue 2: Evaluation of Monolingual Competence of First Language (L1) learners (more niche, but important for some tasks)
- ▶ Towards enhanced Explanatory Adequacy: Understanding Linguistic Mechanisms in LLMs (Causality and Linguistic Interpretability)

String Probability and Grammaticality

“What Can String Probability Tell Us About Grammaticality?”

- ▶ recent paper by Hu et al. (2025)
- ▶ Chomskyan arguments suggests probability and grammaticality are distinct.
- ▶ Hu et al. suggest that the probability of a string (s) depends on message (m) and grammaticality (g)

$$P(\mathbf{s}) = \sum_{\substack{m \in \mathcal{M}, \\ g \in \{0,1\}}} P(\mathbf{s}|m, g)P(g|m)P(m) \quad (2)$$

String Probability and Grammaticality: 3 Predictions

Description	Status
Corr. between the log-prob. of grammatical and ungrammatical strings within minimal pairs	✓
Corr. between log-prob. and acceptability judgments	Mixed
Potentially(?) poor separation between grammatical / ungrammatical strings.	✓

Table: 3 Predictions and results from Hu et al. (2025).

Code-Switching (Sterner & Teufel 2025)

Code-Switching

How do we evaluate the code-switching capabilities of LLMs?

1000 minimal pairs for 11 language pairs of a **naturally occurring CS sentence** (consistently preferred by bilinguals) and a **minimally manipulated variant**.

Code-Switching (Sterner & Teufel 2025)

- (1) a. @USER **And I said maybe** etwas leiser singen, sonst ruf ich die Polizei
b. @USER **And I said maybe a little** leiser singen, sonst ruf ich die Polizei

[German–English, a. from Sterner and Teufel, 2023]

- (2) a. **Would do it myself** om inte make var bilmek och nördig med det.
b. **Would do it själv** om inte make var bilmek och nördig med det.

[Swedish–English, a. from DeLucia et al., 2022]

- (3) a. 同学们 会 大概 知道 **what they want to do in the future**
b. 同学们 会 大概 know **what they want to do in the future**

[Chinese–English, a. from Lyu et al., 2010]

Second Language Acquisition (SLA): BLiSS 1.0

For models that aim to be cognitively plausible, we need a complementary, acquisition-focused perspective, one that inspects how grammar competence is organised and learned. This evaluation gap is important for models of Second Language Acquisition (SLA), which we refer to as L2LMs (Aoyama and Schneider, 2024).

Second Language Acquisition (SLA): BLiSS 1.0

A central characteristic of the SLA process is the production of **systematic errors**. These deviations are not random noise, but rather structured evidence of the learner's developing internal grammar, or **interlanguage**. This is the motivation for **BLiSS 1.0** (Gao, Salhan, Caines, Buttery & Sun 2025)

Second Language Acquisition (SLA): BLiSS 1.0

```
1 {  
2   "learnerID": "8421",  
3   "L1": "Vietnamese",  
4   "cefr": "C1",  
5   "topic": "play sports",  
6   "corrected": "There are a lot of  
   benefits when we play sports.",  
7   "learner error": "There are a lot of  
   benefits when we play the  
   sports.",  
8   "artificial error": "There are a lot  
   of benefits when the we play  
   sports.",  
9  
10  "errant_edits": [{  
11    "type": "U:DET",  
12    "o_str": "the",  
13    "c_str": ""  
14  }],  
15  "all_error_types": [  
16    "U:DET"  
17  ]  
18 }
```

Figure 1: An example BLiSS triplet illustrating an Unnecessary Determiner (U:DET) error. The original learner sentence contains an unnecessary determiner “the”, which is removed in the corrected sentence. Artificially-generated errors of the same type allow controlled evaluation of model preferences.

Searching for Syntactic Structure beyond English. Two lines of current research extend earlier syntactic probing papers:

- ▶ Mechanisms within and across architectures
- ▶ Causality and Circuits: identifying neurons and representations (e.g., comparing multilingual and monolingual representations)

SyntaxGym and CausalGym

Phenomena: **Agreement, Licensing, Garden Path Effects, Gross Syntactic State and Long-Distance Agreement.**

Why? These are phenomena testable for **targeted syntactic evaluation** to learn more about **generalisation** in Language Models.

Linguistic Interpretability needs **systematic minimal pairs that vary by a specific feature** $[F_i]$ (not BLiMP!).

SyntaxGym and CausalGym

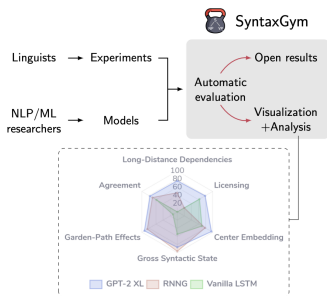


Figure 1: SyntaxGym allows linguists to easily design and run controlled experiments on the syntactic knowledge of language models, and allows NLP experts to test their own models against these standards. Users submit targeted syntactic evaluation experiments to the site, and they are automatically evaluated on language models available in the Gym. SyntaxGym analyzes and visualizes these evaluation results.

Figure from SyntaxGym Paper.

SyntaxGym and CausalGym

We can use these to understand linguistic mechanisms in model families.

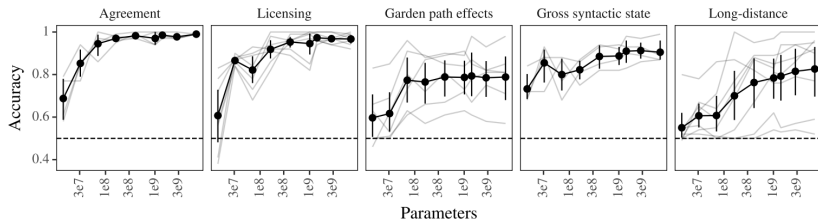


Figure 3: Accuracy of pythia-family models on the **CausalGym** tasks, grouped by type, with scale. The dashed line is random-chance accuracy (50%).

A Brief Aside: Mechanistic Interpretability

- ▶ **Circuit Localization:** methods to locate the most important subsets of a model for performing a given task.
- ▶ **Causal Variable Localization:** Featurising hidden vectors and selecting features that map

These have natural analogues for Linguistic Interpretability research and can be viewed as a methodological extension of earlier Targeted Syntactic Evaluation research.

Interpreting Minimal Pairs in LLMs

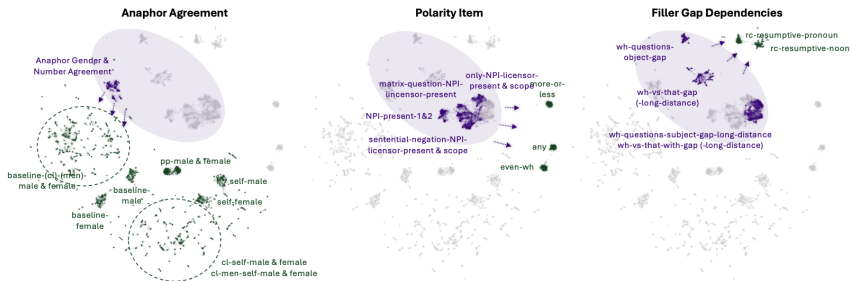


Figure 7: **UMAP visualization of minimal pairs in same categories (terms) but different languages.** Linguistic phenomena are dominantly grouped by their language in multilingual LLMs: English samples are all clustered within purple-shaded areas. While relevant linguistic phenomena in different languages are not fully overlapped, LLM does capture some relationships. As indicated by purple arrows, English samples seem to be “attracted” by the corresponding Chinese samples.

Curse of Bilinguality

Across 55 test tasks, there are consistent performance differences between monolingual and bilingual models on 16 tasks. Despite their smaller sizes, monolingual models perform better on 12 and worse only on 4 tasks. (Zhou & Matushevych, 2025)

Curse of Bilinguality

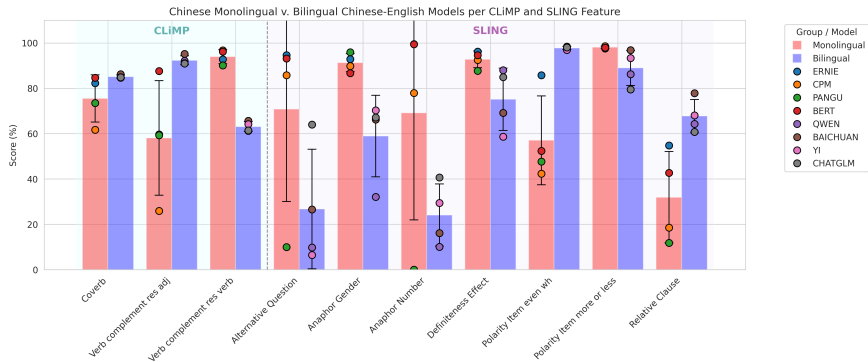


Figure from Suchir Salhan

Curse of Bilinguality

This result suggests that bilingual Chinese + English bilingual models may **suffer from negative crosslingual transfer**. →
Implications for Multilingual Pretraining and Cross-Lingual Transfer

Thanks to Laura Barbenel for finding this paper!

Further Reading: Multilingual Benchmarking

MultiBLiMP 1.0

Paper: <https://arxiv.org/pdf/2504.02768>

Datasets on HuggingFace:

<https://huggingface.co/datasets/jumelet/multiblimp>

BabyBabelLM (Multilingual BabyLM):

<https://arxiv.org/pdf/2510.10159>

Further Reading: Linguistic Interpretability

BLiSS 1.0 (BabyLM 2025): <https://arxiv.org/pdf/2510.19419>

Curse of Bilinguality (COLING 2025):

<https://aclanthology.org/2025.gem-1.58.pdf>

Minimal Pair-Based Evaluation of Code-Switching (ACL 2025):

<https://aclanthology.org/2025.acl-long.910.pdf>

Further Reading: Linguistic Interpretability

SyntaxGym (ACL Systems Demo 2020):

<https://aclanthology.org/2020.acl-demos.10.pdf>

CausalGym (ACL 2024):

<https://aclanthology.org/2024.acl-long.785.pdf>

Next Time: The Trilogy of Chinese BLiMPs

- ▶ CLiMP
- ▶ SLING
- ▶ And, more recently, **ZhoBLiMP** (2024, under review):
<https://arxiv.org/abs/2411.06096>

Thank you for your attention!

References I

- Hu, J., Wilcox, E. G., Song, S., Mahowald, K., & Levy, R. P. (2025, October 17). *What Can String Probability Tell Us About Grammaticality?* arXiv: 2510.16227 [cs].
<https://doi.org/10.48550/arXiv.2510.16227>
- Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 377–392. https://doi.org/10.1162/tacl_a.00321