UNIVERSITY OF CAMBRIDGE

Evaluating the Cross-Lingual Syntactic Capabilities of Language Models

Suchir Salhan, David Strohmaier
{sas245,ds858}@cam.ac.uk

Computer Laboratory, University of Cambridge, UK

L95 2025/26 **27 October 2025**

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

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

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	*Many girls insulted themselves.*	*Many girls insulted herself.*
Arg. structure	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
Binding	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
Control/raising	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
Det.-noun agr.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
Ellipsis	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
Filler-gap	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
Irregular forms	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
Island effects	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI licensing	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
Quantifiers	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
Subject-verb agr.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

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

Binding

Referential Expressions

- Anaphors, i.e. reflexive pronouns such as themself
- 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

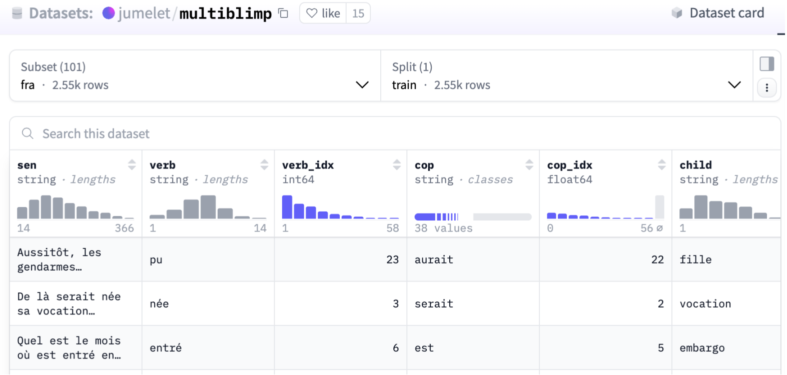
Minimal Pair Datasets Beyond English

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!



Datasets: jumelet/multiblimp like 15 Dataset card

Subset (101)
fra · 2.55k rows

Split (1)
train · 2.55k rows

Search this dataset

sen	verb	verb_idx	cop	cop_idx	child
string · *lengths*	string · *lengths*	int64	string · *classes*	float64	string · *lengths*
14 366	1 14	1 58	38 values	0 56 ⌀	1
Aussitôt, les gendarmes…	pu	23	aurait	22	fille
De là serait née sa vocation…	née	3	serait	2	vocation
Quel est le mois où est entré en…	entré	6	est	5	embargo

🤗 **MultiBLiMP 1.0** on HuggingFace

MultiBLiMP 1.0: Data Construction

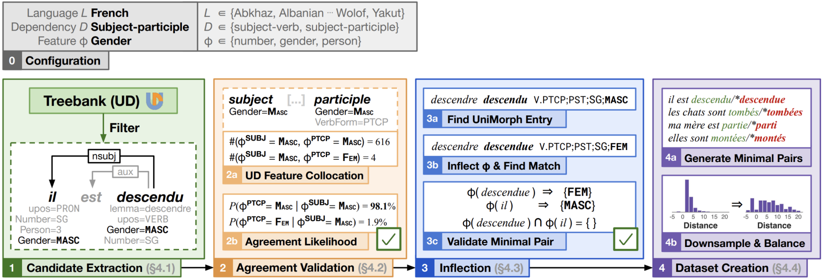

Language L French
Dependency D Subject-participle
Feature ϕ Gender
L ∈ {Abkhaz, Albanian ··· Wolof, Yakut}
D ∈ {subject-verb, subject-participle}
ϕ ∈ {number, gender, person}
0 Configuration
Treebank (UD)
Filter
nsubj
aux
il est descendu
upos=PRON
Number=SG
Person=3
Gender=MASC
lemma=descendre
upos=VERB
Gender=MASC
Number=SG
1 Candidate Extraction (§4.1)
subject [...] participle
Gender=Masc
Gender=Masc
VerbForm=PTCP
#(ϕ^SUBJ = Masc, ϕ^PTCP = Masc) = 616
#(ϕ^SUBJ = Masc, ϕ^PTCP = Fem) = 4
2a UD Feature Collocation
P(ϕ^PTCP = Masc | ϕ^SUBJ = Masc) = 98.1%
P(ϕ^PTCP = Fem | ϕ^SUBJ = Masc) = 1.9%
2b Agreement Likelihood
✓
2 Agreement Validation (§4.2)
descendre descendu V.PTCP;PST;SG;MASC
3a Find UniMorph Entry
descendre descendue V.PTCP;PST;SG;FEM
3b Inflect ϕ & Find Match
ϕ(descendue) ⇒ {FEM}
ϕ(il) ⇒ {MASC}
ϕ(descendue) ∩ ϕ(il) = { }
3c Validate Minimal Pair
✓
3 Inflection (§4.3)
il est descendu/*descendue
les chats sont tombés/*tombées
ma mère est partie/*parti
elles sont montées/*montés
4a Generate Minimal Pairs
-5 0 5 10 15 20
Distance
⇒
-5 0 5 10 15 20
Distance
4b Downsample & Balance
4 Dataset Creation (§4.4)

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 Gwyneed.

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 (**gradience** 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)

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

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 occuring 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]

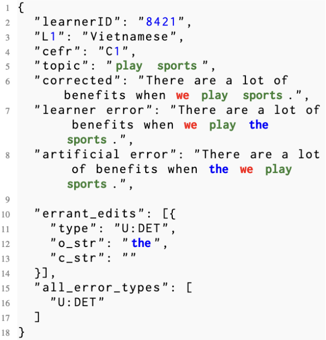
Second Language Acquisition (SLA): BLiSS 1.0

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 learnerâs 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



```
{
  "learnerID": "8421",
  "L1": "Vietnamese",
  "cefr": "C1",
  "topic": "play sports ",
  "corrected": "There are a lot of
      benefits when we play sports .",
  "learner error": "There are a lot of
       benefits when we play the
      sports .",
  "artificial error": "There are a lot
       of benefits when the we play
      sports .",

  "errant_edits": [{
    "type": "U:DET",
    "o_str": "the ",
    "c_str": ""
  }],
  "all_error_types": [
    "U:DET"
  ]
}
```

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.

BLiSS 1.0 (Gao, Salhan, Caines, Buttery & Sun 2025)

Linguistic Interpretability

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

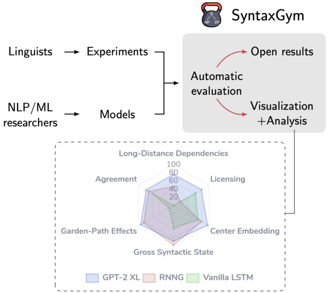


Figure 1: SyntaxGym allows linguists to easily design and run controlled experiments on the syntactic knowledge of language models, and allows NLP experts to test their own models against these standards. Users submit targeted syntactic evaluation experiments to the site, and they are automatically evaluated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

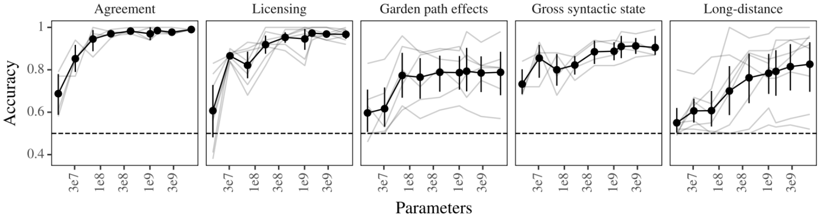


Figure 3: Accuracy of `pythia`-family models on the **`CausalGym`** tasks, grouped by type, with scale. The dashed line is random-chance accuracy (50%).

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

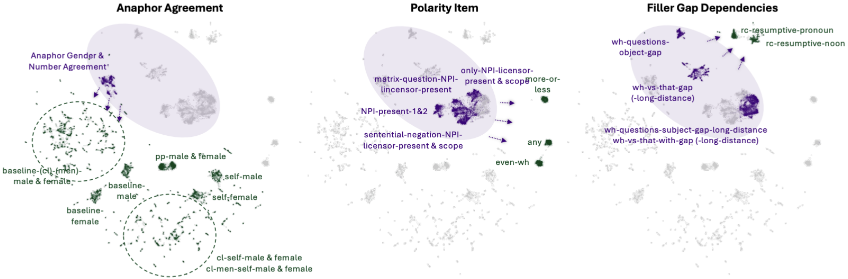


Figure 7: **UMAP visualization of minimal pairs in same categories (terms) but different languages**. Linguistic phenomena are dominantly grouped by their language in multilingual LLMs: English samples are all clustered within purple-shaded areas. While relevant linguistic phenomena in different languages 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 & Matusevych, 2025)

Curse of Bilinguality

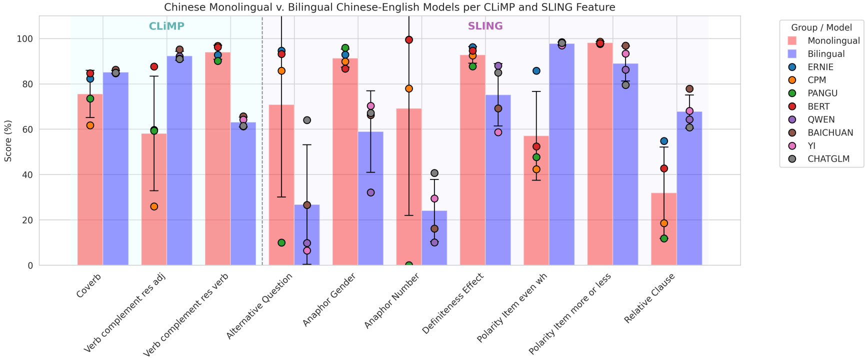


Figure from Suchir Salhan

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