Evaluating the Cross-Lingual Syntactic Capabilities of Language Models

Suchir Salhan

December 30, 2024

We focus on how syntactic theory can be used to evaluate Language Models, and how syntactic typology can be leveraged to develop evaluation metrics for language models in a cross-lingual setting.

Evaluating Grammaticality in Language Models

Language Models can be evaluated intrinsically on datasets that assess different linguistic capabilities or extrinsically on real-world tasks or applications. To assess a model's syntactic competence, we require an appropriate *metric* which can be used to meaningfully compare systems. The linguistic capabilities of language models can be intrinsically evaluated using minimal pairs of datasets consisting of pairs of contrasting grammatical and ungrammatical sentences. This is the dominant method for evaluating Natural Language Syntax. A common approach to arrive at an overall score of the syntactic capabilities of a Language Model is to macro-average the accuracies across test sets covering various syntactic phenomena. [1] Accuracy calculations differ between **causal/autoregressive language models** (e.g., GPT or LLama), where the chain rule is applied by summing the log-likelihood values for each successive token, and **Masked Language Models (MLM)** (e.g., BERT or RoBERTa).

Sentence pseudo-log-likelihood (PLL) scores are estimated for MLMs by successively masking each sentence token, retrieving its score using the rest of the sentence as context, and summing the resulting values. [2]

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

This is based on an interpretation of the MLM objective as a stochastic *maximum pseudolikelihood estimation* (MPLE) on a training set $\mathcal{W}$, which approximates the conventional Maximum Likelihood Estimation (MLE). This is by asymptotically maximising an objective:

$$\mathcal{J}_{\mathrm{PL}}(\Theta; \mathcal{W}) = \frac{1}{|\mathcal{W}|} \sum_{\mathbf{W} \in \mathcal{W}} \mathrm{PLL}(\mathbf{W}; \Theta).$$

, where $\{\mathbf{w}_t\}_{t=1}^{|\mathbf{W}|}$ are random variables in a fully connected graph. In this way, MLMs learn an underlying joint distribution whose conditional distributions $\mathbf{w}_t \mid \mathbf{W}_{\setminus t}$ are modelled by masking at position t.

[1] ***Is Macro-Averaging a meaningful or cognitively plausible way to assess the capabilities of a system?*** Language learners acquire syntactic phenomena concurrently, so macro-averaging may not align with realistic scenarios for evaluating the *development* of linguistic capabilities.

[2] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.240. URL https://aclanthology.org/2020.acl-main.240

This PLL metric is very popular in LLM work that assesses the effect of training data and model fluency. [3]

However, PLL leads to over-inflated scores for out-of-vocabulary (OOV) tokens that are tokenised into subword tokens, which are predicted using a token's bidirectional context. To address this, a new metric `PLL-word-l2r` places a [MASK] over the current target token (now: s_{w_t}), but also over all future sentence tokens that belong to the same word s_w as the target. As shown in *Figure* 1, this is an intermediate strategy to compute a PPL score for an OOV token like `souvenir`, which is tokenised as subwords *so ##uven ##ir* instead of whole word masking and the default strategy. Inference is then conditioned on a context that includes all preceding sentence tokens (including those belonging to the current word) and all sentence tokens from future words. [4] The final score of a sentence S is obtained as the sum of the log probabilities of each of the w tokens in each of the S words:

$$\mathrm{PLL}_{\mathrm{l2r}}(S) := \sum_{w=1}^{|S|} \sum_{t=1}^{|w|} \log P_{\mathrm{MLM}}\left(s_{w_t} \mid S \setminus \{s_{w_{t'\geq t}}\}\right) \tag{1}$$

For both causal and masked language models, probabilities are normalized by sentence length. Recently, an alternative approach has been to prompt LLMs to **rate item plausibility**, absolutely or on a Likert scale.[5] LLMs perform worse with direct prompting and metalinguistic prompts, which cannot be taken as conclusive evidence that LLM lacks a particular linguistic generalisation [6]. A new evaluation metric called the **Elements of World Knowledge (EWoK)**.[7] uses both traditional plausibility estimates via log probability and two prompt-based strategies called LIKERT and CHOICE. The metric for correctness of a given item is the recovery of the designed item structure such that

$$\mathrm{score}(T_1 \mid C_1) > \mathrm{score}(T_1 \mid C_2)$$

and

$$\mathrm{score}(T_2 \mid C_1) < \mathrm{score}(T_2 \mid C_2),$$

where score reflects P_θ for log probabilities, an integer rating for LIKERT, and the correct context index selection for CHOICE, and T is the target sentence and C is the context of the minimal pair.

Another possibility is evaluating models on the probability that a language model assigns to a **critical word**, which is the word in the sentence where it can become ungrammatical. Dubbed the *"two-prefix method"*, we would expect the language models to give this word in particular a lower probability in the ungrammatical than in the grammatical sentences. As the critical word will be the same

[3] Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online, August 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.90. URL `https://aclanthology.org/2021.acl-long.90`

[4] Carina Kauf and Anna Ivanova. A better way to do masked language model scoring. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada, July 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.acl-short.80. URL `https://aclanthology.org/2023.acl-short.80`

[5] The latter is standard experimental practice in Experimental Syntax. If you are interested about best practices in acceptability experiments, consider reading the introductory chapters in *The Cambridge Handbook of Experimental Syntax*

[6] Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore, December 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.306. URL `https://aclanthology.org/2023.emnlp-main.306`

[7] Here is some more information about this: `https://ewok-core.github.io/`

for ungrammatical/grammatical sentences and the frequency of the critical word is the same, the only thing that differs is the **preceding context**.

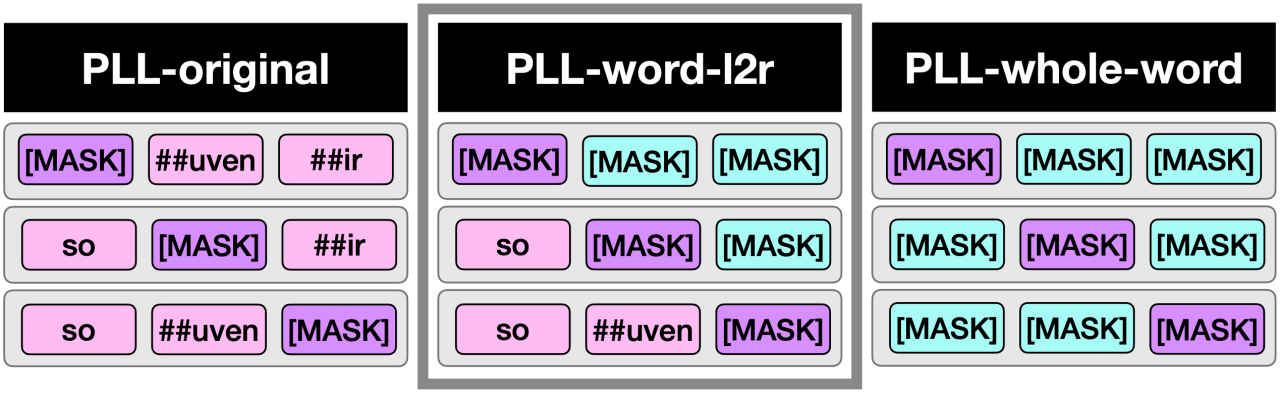

Figure 1: The PLL score of a multi-token OOV items, split into subword tokens, can be computed in different ways. *Purple*: target token, *pink*: within-word tokens that are available during inference, *turquoise*: within-word tokens that are masked during inference. Sentence tokens that do not belong to the current word are always available during inference. *Figure from Kauf & Ivanova (2023)*

Linguistic Limitations:

- Scores can be influenced by superfluous factors, e.g., the number of available synonyms. Therefore, PLL approaches are only useful in highly restricted minimal pair setups
- A revised PLL_{l2r} metric may not generalise to agglutinative languages (increased uncertainty due to a high number of tokens per word) and the metric probably is not as important for isolating/analytic languages where word-level items can be represented as single tokens.

BLiMP: Benchmark of Linguistic Minimal Pairs

BLiMP consists of 12 syntactic phenomena in English with unique identifiers (UIDs) of 67 syntactic paradigms. NLP practitioners standardly report BLiMP macro-averages.[8]

- Minimal pairs were **artificially generated** using from abstract grammars that exemplify syntactic phenomena – this easily yields a large number of sentences, which can help control for other possible sources of noise in test materials. Generation scripts use templates to sample lexical items with **selectional restrictions**, which annotate the morphological, syntactic, and semantic features of over 3000 items.
- **Human Evaluation:** Human benchmarking is important in several NLP tasks. It is a useful proxy for the difficulty of different tasks. For BLiMP, the authors used 20 validators who rated five pairs from each of the 67 paradigms for 6,700 judgments.

[8] BLiMP: The Benchmark of Linguistic Minimal Pairs for English, available at: `https://aclanthology.org/2020.tacl-1.25.pdf`

BLiMP is standardly used for evaluating monolingual English Language Models with other semantic evaluation benchmarks like the **General Language Understanding Evaluation (GLUE) benchmark**.[9] *Table* 1 shows an example of benchmarking in the BabyLM Shared Task, which uses BLiMP alongside GLUE and EWoK.

[9] Here is the GLUE paper: `https://openreview.net/pdf?id=rJ4km2R5t7`

Model	BLiMP	BLiMP Suppl.	EWoK	GLUE	Av.
BabyLlama	69.8	59.5	50.7	63.3	60.8
LTG-BERT	60.6	60.8	48.9	60.3	57.7

Table 1: Example of Language Model Evaluation from the BabyLM Shared Task 2024

Agreement

BLiMP's Subject-verb agreement dataset has minimal pairs of contrast sentences with correct and incorrect agreement (e.g., *These casseroles **disgust**/***disgusts** Kayla*). Similarly, BLiMP's Determiner-noun agreement dataset consists of minimal pairs about number agreement between demonstrative determiners (e.g. *this*/*these*) and the associated noun. The determiner-noun agreement and subject-verb agreement phenomena also include paradigms illustrating irregular morphology. BLiMP's Irregular forms contain irregular English past participles morphology in an adjectival case (*The forgotten newspaper article was bad.* v **The forgot newspaper article was bad.*) and a verbal case (*Edward hid the cats.* v *Edward hidden the cats.*) BLiMP does not evaluate models on non-existent forms like **breaked* because such forms are out of the vocabulary for some LMs.

Filler-Gap Dependencies and Island Effects

Filler-Gap Dependencies arise from phrasal movement in, e.g., *wh*-questions. BLiMP's dataset contains minimal pairs across *interveners*. These include **subject gaps** (e.g., *Cheryl thought about **some dog** that upset Sandra* v. **Cheryl thought about who some dog upset Sandra.*) and **object gaps** (e.g., *Joel discovered the vase that Patricia might take.* v **Joel discovered what Patricia might take the vase.*). Filler-gap dependencies can be **long-distance dependencies** with interveners:
Susan won't discover a car that Jane admired that is aggravating a lot of cashiers./Susan won't discover who a car that Jane admired is aggravating a lot of cashiers. **(subject gap)**
*Laurie forgot some alumnus that most organizations that competed have known./ *Laurie forgot who most organizations that competed have known some alumnus.* **(object gap)**

Figure 2 shows that GPT 2 and GPT 3 have a negative wh-effect in the gap condition and a positive wh-effect in thegap condition, which

shows that to some degree models are learning the basic filler-gap dependency.

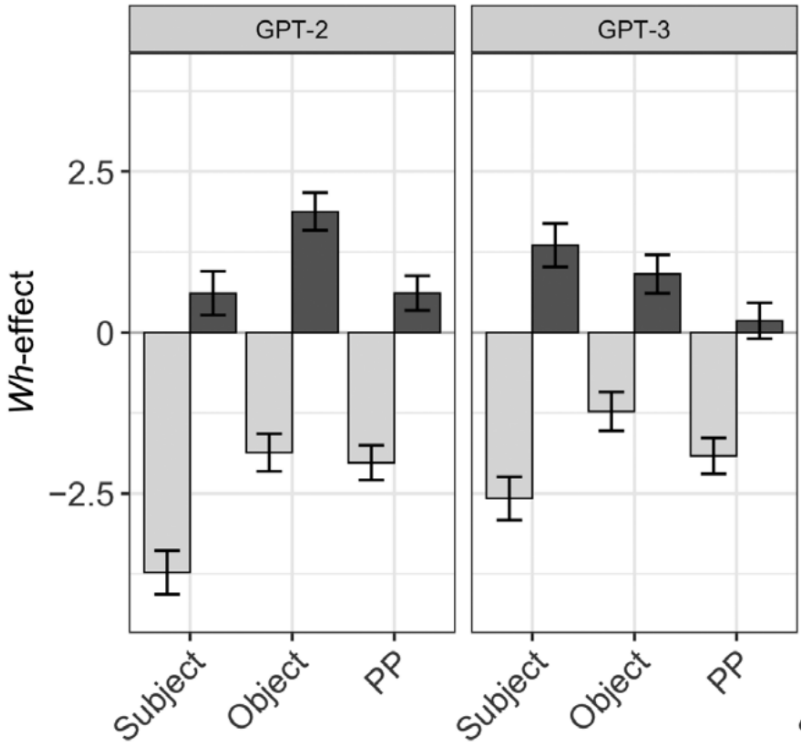

Figure 2: GPT-2 and GPT-3 show sensitivity to island conditions. *Figure from Wilcox, Futrell & Levy (2024) "Using Syntactic Models to Test Syntactic Learnability"*, available from: `https://www.colinphillips.net/wp-content/uploads/2024/05/wilcox2023.pdf`

Island Effects characterise restrictions on syntactic environments where the gap in a filler-gap dependency may occur. Descriptively, we can identify several classes of ungrammatical sentences summarised in *Table* 4, where the strings written in brackets indicate "copies" (or traces, in earlier terminology) of constituents under syntactic theories assuming movement.

- Adjunct Islands: Gaps cannot be licensed inside an adjunct clause
- Complex NP Islands: Gaps are not licensed inside S nodes that are dominated by a lexical head noun
- Coordination Islands: Gaps cannot occur in only one half of a coordinate structure
- Left Branch Islands: Modifiers that appear on the "left branch" under an NP cannot be gapped.

Statistical evidence for island effects have been found across Language Models by contrasting the wh-effects in an island condition

a. *Adjunct islands*
   *I know what the patron got mad after the librarian placed ___ on the wrong shelf.
b. *Complex NP islands*
   *I know what the actress bought the painting that depicted ___ yesterday.
c. *Coordinate structure islands*
   *I know what the man bought and ___ at the antique shop.
d. *Left-branch islands*
   *I know how expensive you bought ___ a car last week.
e. *Sentential subject islands*
   *I know who for the seniors to defeat ___ will be trivial.
f. *Subject islands*
   *I know who the painting by ___ fetched a high price.
g. Wh-*islands*
   *I know who Alex said whether your friend insulted ___ yesterday.

Figure 3: Islands associated with syntactic constraints, based on Ross (1967) and Huang (1982)

with the wh-effects in a nonisland minimal-pair counterpart, typically with gaps in object position. GPT-2 was found to be sensitive to all islands. Despite cross-model architectural variation, strongest effects have been found for coordination, adjunct, and complex NP islands. However, language models have weaker effects for left-branch, subject, and sentential subject islands [10].

[10] Ethan Gotlieb Wilcox, Richard Futrell, Roger Levy; Using Computational Models to Test Syntactic Learnability. Linguistic Inquiry 2024; 55 (4): 805–848. https://www.colinphillips.net/wp-content/uploads/2024/05/wilcox2023.pdf

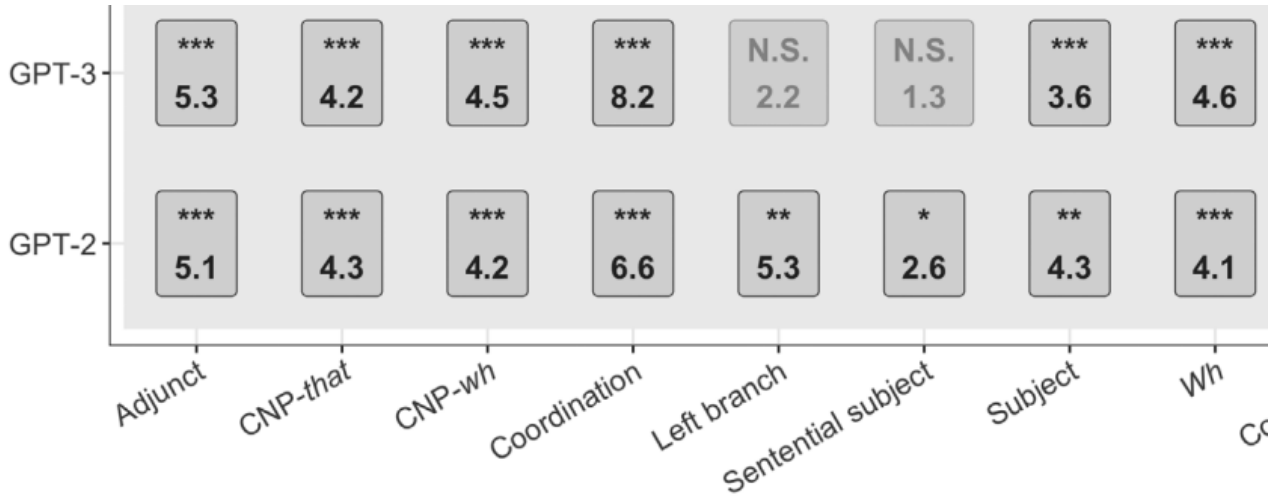

Figure 4: Summary of wh-effects across island test sets for GPT-2 and GPT-3

Binding

Syntacticians distinguish **anaphors** (reflexive pronouns like him/her/themselves), **pronouns** (he, her) and **R-expressions**, which are Noun Phrases that get meaning from referring to an entity in the world. BLiMP's ANAPHOR AGREEMENT dataset contains minimal pairs that differ in the grammaticality of anaphors, which are required to agree with their antecedents in person, number, gender and animacy.

BLiMP's BINDING dataset contains properties of the structural relationship between a pronoun and its antecedent. All paradigms, summarised in *Table* , illustrate aspects of Principle A, which characterises restrictions on the distribution of anaphors.

Category	**Sentence**
C-command	A lot of patients who can sell some couch didn't investigate **themselves/*itself**.
Principle A Case 1	The teenagers explain that **they/*themselves** aren't breaking all glasses.
Principle A Case 2	Eric imagines himself **taking/*took** every rug.
Domain 1	Carla had explained that Samuel has discussed **her/*herself**.
Domain 2	Donald can imagine those college campuses are boring **themselves/*himself**.
Domain 3	Steven explains Kayla won't hurt herself v Kayla explains Steven won't hurt herself.
Reconstruction	It's himself that **this cashier attacked/*attacked this cashier**.

Since co-indexation cannot be annotated in BLiMP, Principles B and C, which characterise restrictions on pronouns and R-expressions, are not contained in the minimal pairs dataset. Binding Principle B precludes pronouns from being locally bound in the same manner as

anaphors (e.g., *Mary said that Joe liked these pictures of her* v **Mary said that Joe liked these pictures of him*).

Control and Raising

BLiMP's CONTROL/RAISING constructions highlight syntactic and semantic differences between various types of predicates in non-finite clauses which embed an infinitival VP. We can broadly identify two types of constructions that lack an overt subject.

Raising constructions have a predicate with a syntactic argument that is naturally the semantic argument of its embedded predicate. In a sentence like *He seems to scare them.*, *seem* does not "select" the subject. Assuming a syntactic theory with movement, the argument *he* moves from the embedded clause to its subject position. Meanwhile, in **control** constructions, the matrix verb "controls" the arguments in the subordinate clause, e.g., in the sentence *John promises to help us*, the subject *John* is the controller of the arguments in *help* clause. We refer to *promise* as a control verb, which semantically selects its arguments.[11]

[11] The syntax of control varies across formalisms. Generative linguists posit a null element PRO to formally accomodate control constructions in X-bar theoretic analyses

Note that these dependencies are not standardly represented in *basic* Universal Dependencies. **Enhanced Universal Dependencies Graphs** represents control and raising constructions via an **additional dependency** (i.e. an additional `nsubj`) between a controlled verb and its controller or between an embedded verb and its raised subject.

BLiMP's datasets contain three types of raising and control constructions:

- ***tough*-movement predicates:** These are predicates involving verbs like *tough/difficult/easy* that allow the subject of the matrix clause to appear semantically as the object of the embedded clause. An example of this contrast in the BLiMP dataset is: *Julia wasn't fun to talk to.* v **Julia wasn't unlikely to talk to*
- **Existential *there*:** *there* is used to indicate that something exists, or to assert its non-existence. *William has declared there to be no guests getting fired.* v **William has obliged there to be no guests getting fired.*
- **Expletive *it*:** Dummy *it* is introduced in cases of raising (and extraposition). *Carla could declare it to be not so important that these doctors observe Rhonda.* v **Carla could convince it to be not so important that these doctors observe Rhonda.*

The syntax of control and raising extends beyond these simple cases of subject raising and subject control. English can have **object raising** and **object control**.[12] These cases are not explicitly handled in BLiMP.

[12] Raising to Object verbs are also known as Exceptional Case Marking (ECM) Verbs. These are infinitives that have embedded accusative subjects, e.g., *Rosie believed him to be innocent.*

Ellipsis

BLiMP does not offer full coverage of ellipsis, since it only considers sentences of equal length. The ellipsis paradigms cover special cases of NP ellipsis (or more, precisely, in X-bar terms **N-bar Ellipsis**) that meet this practical constraint:
Brad passed one big museum and Eva passed several. v * *Brad passed one museum and Eva passed several big*.

It is worth mentioning that English has several forms of **predicate/VP ellipsis (VPE)**:

- Auxiliary VPE: Susan has read War and Peace, but Maria hasn't.
- Modifier VPE: Susan can speak French, and Maria can too.
- Pseudogapping: Susan doesn't eat pasta, but she does pizza.
- Antecedent Contained Deletion: Susan has read every book Maria has.

Typologically, we can note that many Romance and Germanic languages lack Auxiliary VPE, although they do have Auxiliary VPE, and pseudogapping is also more marginal here. Syntacticians typically attribute these differences to the nature of the English auxiliary system.

Syntax-Semantic Interface

BLiMP contains three "interface" phenomena:

- ARGUMENT STRUCTURE: the ability of different verbs to appear with different types of arguments. BLIMP's Argument Structure consists of verbs that appear with a direct object, participate in a **causative alternation** (*the boy broke the window* v *the window broke*), or take an inanimate argument.
- NPI LICENSING: restrictions on the distribution of **negative polarity items** like *any* and *ever*. limited to, e.g., the scope of negation and *only*.
- QUANTIFIERS: restrictions on the distribution of quantifiers. We cover two such restrictions: superlative quantifiers (e.g., *at least*) cannot embed under negation, and definite quantifiers and determiners cannot be subjects in existential-*there* constructions.

BLiMP Supplement

BLiMP Supplement was unofficially released for the BabyLM Shared Task and additionally contains minimal pairs datasets for **subject-auxiliary inversion** (e.g., *Is the novel he is putting away from the library?*

v *Is the novel he putting away is from the library?*), and **hypernyms** (e.g., *If she has a dog, it must be the case that she has a mammal.* v *If she has a dog, it must be the case that she has a chihuahua.* Additionally, it contains **discourse phenomena** like **turn taking**:
"David: Should you quit? Sarah: No, I shouldn't."
*? *"David: Should she quit? Sarah: No, I shouldn't."*

Additionally, it contains datasets about **question-answering congruence**. This is the only minimal pairs dataset that is split into **difficulty levels** (EASY/HARD). An inimate v animate contrast is meant to be *easy*, while an animate vs. inanimate is meant to be *tricky*.
Easy: *"What did you get? I got a chair."* v *? *"What did you get? I got a teacher."*
Tricky: *"Who cleaned? David cleaned."* v *? *"Who cleaned? The patio cleaned."*

Syntactic Typology

Minimal Pairs datasets have been introduced beyond English:

1. **CLAMS (French and German):** The Cross-Lingual Syntactic Evaluation of Word Prediction Models (CLAMS) [13] generates minimal pair datasets which we use for French and German using Attribute-Varying Grammars. The dataset assesses grammaticality in Simple Agreement, VP coordination, and across "interveners" in S-V agreement (subject/object relative clause or across a Prepositional Phrase).
2. **JBLIMP (Japanese):** JBLIMP [14] is a minimal pairs dataset for targeted syntactic evaluation of Japanese. It consists of 331 minimal pairs of syntactic acceptability judgements curated from Japanese syntax articles in the *Journal of East Asian Linguistics*. The JBLiMP Minimal Pair dataset can be found here: `https://github.com/osekilab/JBLiMP/tree/main`
3. **SLING (Chinese): SLING** [15] is a 38K minimal sentence pair dataset derived by applying syntactic and lexical transformations to Chinese Treebank 9.0, aiming to improve on the limitations of an earlier dataset called CLiMP [16], which had a lack of diversity in the vocabulary to generate minimal pair templates. The SLING Dataset can be found here: `https://huggingface.co/datasets/suchirsalhan/SLING`

Due to the small size of the JBLIMP minimal pairs dataset, Someya and Oseki [2023]'s recommend to compute accuracy using a SLOR score to mitigate the confounding effects of lexical frequencies and sentence lengths, which is defined as follows:

[13] Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.490. URL `https://aclanthology.org/2020.acl-main.490`

[14] Taiga Someya and Yohei Oseki. JBLiMP: Japanese benchmark of linguistic minimal pairs. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-eacl.117. URL `https://aclanthology.org/2023.findings-eacl.117`

[15] Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. SLING: Sino linguistic evaluation of large language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. DOI: 10.18653/v1/2022.emnlp-main.305. URL `https://aclanthology.org/2022.emnlp-main.305`

[16] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the*

$$SLOR(X) = \frac{\log p_m(X) - \log p_u(X)}{|X|}$$

where $p_m(X)$ is the probability of a sentence for a Language Model and is the unigram probability of the sentence, estimated for each subword in the training corpus. Accuracy calculations for other languages follows dataset guidance to use unnormalised log-probabilities.

BLiMP, CLiMP, SLING and JBLiMP all use a forced-choice paradigm to validate their minimal pairs with human native speakers. All papers explore the effect of training data size – CLiMP and JBLiMP found no influence of dataset size. while SLING found that smaller models may have performed better for some. The performance gap between the LMs and the native speakers is large on these cross-lingual minimal pairs datasets (and larger than it was for English). Also, models perform better at local dependencies compared to longer-distance dependencies.

Chinese Syntax: SLING highlights a few important properties of Mandarin syntax. Chinese has a rich system of **classifiers**, so there is an additional syntactic task of **classifier-noun agreement** when a noun is modified by a numeral or demonstrative. **Chinese Definiteness Effect** is a restriction of the distribution of *zhe* (this)/*na* (that) and the quantifier *mei* (every), which may not occur in the post-verbal position of an existential *you* (there is) sentence. Chinese has perfective aspect markers *le* and *guo*. SLING contains minimal pairs that contrast these markers with the tense and the progressive marker *zai*.

Japanese Syntax: Japanese has analytic morphology, so JBLIMP generalises BLiMP's irregular forms dataset to incorporate minimal pairs on morphology in general. Japanese doesn't have explicit determiner-noun agreements, so JBLiMP drops BLiMP's determiner-noun agreement category for a more general Nominal Structure dataset.

BLiMP-NL is a carefully designed new minimal pairs dataset for Dutch, which requires the critical region must be the same for the sentences of the minimal pair, unlike BLiMP to facilitate easier evaluation of langauge models and human evaluation. They source their sentences from Dutch Syntax handbooks.[17] There is also **RuBLiMP**, a Russian BLiMP-style dataset.[18]

Syntactic Typology characterises the dimensions of variation and universality in natural language syntax, for example in word order, alignment, and relative clauses. It is important to be aware that these datasets are (currently) generated from sentences collected by linguists, so are not generated from a perspective that incorporates any notion of **typological distance or granularity**. This inadvertently

[17] https://osf.io/preprints/psyarxiv/mhjbx

[18] https://aclanthology.org/2024.emnlp-main.522.pdf

de-prioritises structures that are cross-linguistically interesting (e.g., ergativity) that are not found in English.

Interpretability Techniques

Individual scores across datasets are not enough to understand the syntactic capabilities of a language model. The Minimal Pairs Paradigm compares the probability of two stand-alone text sequences without any explicit linguistic context. But, this is not necessarily a **naturalistic/realistic** approach, as local contextual information and discourse context can potentially influence grammaticality judgements.[19] Language Models exhibit changes in model performance that are not explainable by acceptability-preserving syntactic perturbations: LLMs have been found to have quantitative, graded effects of structural priming on string probabilities, subject to the length of the context, arising as a consequence of the in-context learning capabilities of Transformer architectures.[20] This has lead to interest in **interpretability techniques** for evaluating the syntactic capabilities of LMs.

[19] *Syntactic Priming* is widely studied in psycholinguistics, studying the effect of linguistic contexts with only one or a small number of context sentences

[20] `https://aclanthology.org/2023.acl-long.333.pdf`

SyntaxGym

SyntaxGym is a syntactic evaluation benchmark designed with more stringent evaluation criteria. For 34 different linguistic phenomena, the SyntaxGym benchmark defines test items with two to four different conditions, consisting of minimal structural variations on the same sentence which render the sentence either grammatical or ungrammatical. Model log-likelihoods are measured at a critical region within each sentence, rather than across the whole sentence, and models are expected to produce log-likelihoods that satisfy multiple inequalities across all conditions.[21]

[21] Hu et al (2020) *A Systematic Assessment of Syntactic Generalization in Neural Language Models*, `https://aclanthology.org/2020.acl-main.158/`. Here are the test sets: `https://syntaxgym.org/`

Syntactic Circuits: A Mechanistic Interpretability Approach

Unlike earlier **probing strategies**, which use a small model trained to extract linguistic information from a target model, causal interventions are the dominant methodology in current mechanistic interpretability work.[22]

Applying this to grammaticality detection, we can adopt a **causal intervention paradigm** to assess grammaticality. The core idea of an intervention is to take a base input b and a source input s and replace a given model-internal component (a "neuron"), f, with $f * (b, s)$, and assess the effect of this intervention on model output to establish causal relationships.

[22] Traditionally, if the probe could predict a target structure, then it was argued that the probe can predict a particular structure, so it is the model it's trained on has implicitly learned to encode it. However, a probe achieving high classification accuracy provides no guarantee that the model actually distinguishes those classes in downstream computations.

CausalGYM takes an input minimal pair that has an alternation that affects next-token prediction, then intervenes on the base forward pass using a pre-defined intervention function that operates on aligned representations from both inputs. Then, it is possible to determine how this intervention impacts next-token prediction probabilities. In aggregate, such interventions assess the causal role of the intervened representation on the model's behaviour.[23]

We can use **directionality** for causal effect is an intuitive test for whether they reflect features that the model uses downstream. **Distributed alignment search (DAS)** learns the **intervention direction**, potentially distributed across many neurons, that maximises the output probability of a **counterfactual label**. The counterfactual label is obtained by recasting a minimal pair, like S-V agreement, from SyntaxGYM into counterfactual pairs that elicit singular or plural verbs based on the number feature of the subject, and hold everything else (including the distractor) constant: (a) *The author near the senators* $\rightarrow$ *is* (b) The authors near the senators $\rightarrow$ are. One of the advantages of this paradigm is that it facilitates an analysis of **model learning dynamics** rather than analysing input/output relationships. Experiments have only been conducted for English and a limited test set, so there is scope for further empirical work in this area. Another line of research identifies **"circuits"** in LMs that handle different tasks. [24]

[23] https://aclanthology.org/2024.acl-long.785/

[24] https://aclanthology.org/2024.findings-emnlp.591.pdf

Beyond BLiMP-style Datasets: Personal Conclusions on Linguistic Evaluation Metrics

Neural Language Models rely heavily on the input that aligns with individual constructions and can even struggle with certain dependencies such as topicalisation. However, where they currently suffer is **generalising *beyond the input*** to learn a shared representation for a given construction. BLiMP and similar datasets cross-lingually do not necessarily characterise various aspects of human syntactic competence. If model evaluation is meant to be theory-agnostic, there is an additional criticism of whether the properties encoded in evaluation datasets are what we should be evaluating (e.g., the top-down desiderata of Construction Grammar may well differ from Generative Grammars),[25].

Additionally, minimal pairs codify an incorrect assumption that there is a strict grammaticality decision boundary. Psycholinguists, for example, have found evidence of **syntactic satiation** – comprehenders, can for example, find island-violating sentences (e.g., **What did John think a bottle of fell on the floor?*) increasingly acceptable given repeated exposure. Syntactic Acceptability is contingent on **linguistic adaptivity** and speaker-specificity; **parsability** is another important

[25] https://aclanthology.org/2024.clasp-1.7.pdf

constraint. These "edge cases" emphasise the gradient nature of acceptability. Ideally, **meta-data** in test sets (e.g., including information about speaker dialects) would be a useful additional source of information, potentially facilitating more fine-grained information about morphosyntactic differences among dialects and varieties.

Overall, there are three main takeaways from a careful analysis of BLiMP. First, it is very important to **scrutinise the datasets you are using off the shelf**. In this case, as is also the case when evaluating models on semantic evaluation datasets, a qualitative evaluation can often tell you more than reporting scores, particularly considering what the limitations of the datasets are (e.g., BINDING does not offer full coverage). Qualitative evaluation of inter-model performance and comparing models to human ceilings can tell you much more about a model than simply reporting a macro-average. Solely reporting macro-averages does not encapsulate any notion of causality in model learning and does not provide a meaningful assessment of learning trajectories in pre-training. Secondly, it is important to be able to assess what the appropriate means of **reporting scores** for a given task is – differences in accuracy calculations are contingent on model architecture and it is important to consider morphological type and frequency beyond English. Finally, like many other areas in NLP, minimal pair datasets suffer from the standard flaws of "starting from English", and subsequent non-English datasets inherit BLiMP's architecture, which may not be ideal. Even beyond English, these datasets have only been built for high-resource languages, and the extent to which the generation of minimal pairs can be feasibly conducted in low-resource regimes is currently unclear.

References

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore, December 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.306. URL https://aclanthology.org/2023.emnlp-main.306.

Carina Kauf and Anna Ivanova. A better way to do masked language model scoring. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada, July 2023. Association for Com-

putational Linguistics. DOI: 10.18653/v1/2023.acl-short.80. URL https://aclanthology.org/2023.acl-short.80.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.490. URL https://aclanthology.org/2020.acl-main.490.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.240. URL https://aclanthology.org/2020.acl-main.240.

Taiga Someya and Yohei Oseki. JBLiMP: Japanese benchmark of linguistic minimal pairs. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-eacl.117. URL https://aclanthology.org/2023.findings-eacl.117.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. SLING: Sino linguistic evaluation of large language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. DOI: 10.18653/v1/2022.emnlp-main.305. URL https://aclanthology.org/2022.emnlp-main.305.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online, April 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.242. URL https://aclanthology.org/2021.eacl-main.242.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online, August 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.90. URL https://aclanthology.org/2021.acl-long.90.