A Typologically Grounded Evaluation Framework for Word Order and Morphology Sensitivity in Multilingual Masked LMs

Anna Feldman, Libby Barak, Jing Peng
Montclair State University
New Jersey, USA
{feldmana, barakl, pengj}@montclair.edu

Abstract

We introduce a typology-aware diagnostic for multilingual masked language models that tests reliance on word order versus inflectional form. Using Universal Dependencies, we apply inference-time perturbations: full token scrambling, content-word scrambling with function words fixed, dependency-based head–dependent swaps, and sentence-level lemma substitution (+L), which lemmatizes both the context and the masked target label. We evaluate mBERT and XLM-R on English, Chinese, German, Spanish, and Russian. Full scrambling drives word-level reconstruction accuracy near zero in all languages; partial and head–dependent perturbations cause smaller but still large drops. +L has little effect in Chinese but substantially lowers accuracy in German/Spanish/Russian, and it does not mitigate the impact of scrambling. Top-5 word accuracy shows the same pattern: under full scrambling, the gold word rarely appears among the five highest-ranked reconstructions. We release code, sampling scripts, and balanced evaluation subsets; Turkish results under strict reconstruction are reported in the appendix.

Keywords: multilingual evaluation, masked language models, morphology

1. Introduction

Multilingual transformer models such as mBERT and XLM-R (Conneau et al., 2020) achieve strong cross-linguistic performance, but what kinds of linguistic cues actually drive their predictions remains unclear. Languages differ sharply in how they encode grammatical relations: fixed-order languages rely on strict sequencing (e.g., English, Chinese), while case-marking or richly inflected languages allow greater flexibility through morphology (e.g., Russian, German, Spanish). If multilingual models depend mainly on linear order, they risk an English-centered bias that disadvantages morphologically rich or free-order languages. If they exploit morphology, they should be more resilient under word-order disruption. Understanding this balance is essential for evaluating the typological generalization and fairness of multilingual encoders.

We treat word order and morphology as partially redundant structural cues and ask how strongly current multilingual encoders rely on each under controlled disruption. The goal is a *diagnostic*: to quantify *relative sensitivity* and *non-additivity* across languages, not to crown a winner between order and morphology.

We do not train new models or argue that reliance on multiple cues is inherently bad. Natural languages routinely exhibit redundancy and multiple exponence. Our contribution is an evaluation protocol and typologically grounded stress tests that expose where models are brittle and where cues overlap.

Prior work shows that transformer-based language models exhibit strong positional biases (Futrell et al., 2019; Warstadt et al., 2020; Papadimitriou et al., 2022), with scrambling often causing sharp performance drops (Ettinger, 2020; Fung et al., 2024). However, most analyses are monolingual and focus on English, leaving open whether multilingual models compensate for disrupted order by drawing on morphology.

We present a typologically grounded evaluation framework for probing the relative roles of word order and morphology in multilingual masked language models. The framework uses Universal Dependencies (UD) treebanks to apply controlled perturbations at inference time: (i) word-order permutations and (ii) sentence-level lemma substitution (+L). We test mBERT and XLM-R on five typologically diverse languages (English, Chinese, German, Spanish, Russian), combining +L with multiple permutation levels. To ensure comparability, masking is applied before any transformation so the same target token is evaluated across all conditions.

Full scrambling collapses accuracy in every language, while partial scrambling yields large but language-dependent drops. Lemma substitution (+L) has strongly language-dependent effects: it is nearly identity in Chinese but substantially lowers accuracy in German/Spanish/Russian. It never compensates for lost order; tiny gains in Chinese are consistent with tokenization/reconstruction quirks rather than genuine morphological leverage. Interaction analyses show that the combined effects are typically sub-additive (joint harm is smaller than additive expectations) indicating

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2154–2165
11–16 May 2026.

overlapping rather than complementary cues. Current multilingual encoders are strongly sensitive to word-order disruption in this diagnostic setting; lemma-normalized contexts (+L) do not yield robustness when order cues are degraded. Because +L is a coarse, sentence-level normalization that changes the label space (surface form vs. lemma), we interpret it as an inflection-stripping diagnostic rather than a feature-specific test of case/agreement/tense.

2. Related Work

Neural language models represent syntactic structure differently depending on architecture and training data. While RNNs struggle with long-range dependencies (Linzen et al., 2016; Gulordava et al., 2018), transformers leverage self-attention mechanisms that, in principle, allow them to capture hierarchical syntax (Vaswani et al., 2017). In practice, however, transformers exhibit strong positional biases, often relying heavily on surface-level word order (Futrell et al., 2019; Warstadt et al., 2020; Papadimitriou et al., 2022).

One line of work probes these biases by scrambling word order either during training or inference. Some studies show that models pretrained on shuffled text can still achieve strong performance, suggesting a surprising degree of robustness to word order disruption (Sinha et al., 2021; Hessel and Schofield, 2021; Gupta and Jaggi, 2021). However, other work finds that performance is highly sensitive to the type of perturbation and the level of linguistic abstraction required by the task (Chen et al., 2024; Papadimitriou et al., 2022). For example, inference-time scrambling often leads to sharp performance drops in tasks requiring syntax-aware reasoning (Pham et al., 2019; O'Connor and Andreas, 2021), though most such work is limited to English or monolingual settings.

Multilingual analyses of word order sensitivity remain rare. Some work investigates scrambling in translation or syntactic transfer (Ahmad et al., 2019; Liu, 2020), while others use token permutation to test cross-linguistic generalization (Zhao et al., 2020). Yang et al. (2019) found that attention-based models are fragile under token swaps, and Ettinger (2020) showed that scrambling impairs masked token prediction even without downstream fine-tuning. Attempts to mitigate this brittleness include introducing de-scrambling objectives (Wang et al., 2019) or auxiliary training on reordered inputs.

Prior probes mostly use local swaps or constrained shuffles (Papadimitriou et al., 2022; Fung et al., 2024), with some exploring stronger/global permutations (Chen et al., 2024); dependency-aware scrambling remains rare. We add a dependency-aware *Head* condition that swaps a UD head with one of its dependents, targeting predicate-argument anchors while keeping other tokens fixed. We use gold UD trees for cross-lingual comparability, though this can still reduce well-formedness.

Our work extends this line of inquiry in four key ways:

1. We test word order sensitivity across five typologically diverse languages with varying morphological richness and syntactic flexibility (see Table 3 for per-language balanced N).
2. We evaluate two widely used multilingual masked language models, mBERT and XLM-R, under identical perturbation settings (results summarized in Tables 4-5).
3. We include a coarse inflection-stripping diagnostic via a sentence-level lemma substitution (+L) that replaces all tokens with their UD `LEMMA` and evaluates the lemma of the masked target; this probes behavior in lemmatized contexts (condition examples in Table 1; corresponding columns in Tables 4-5).
4. We probe graded order disruption with three manipulations: (i) *Full* (all UD word tokens permuted), (ii) *Part* (content words permuted; function words fixed), and (iii) *Head* (each UD head swapped with one of its dependents). (Definitions/examples are in Table 1; results are in Tables 4-5)

By combining structured word order manipulations with lemma normalization across diverse languages, we offer new evidence about how multilingual transformers balance surface structure and grammatical cues in syntactic generalization.

3. Evaluation Framework

We evaluate how word order and surface morphology affect masked language model (MLM) performance using inference-time perturbations applied to two multilingual encoders (mBERT, XLM-R) across five languages: English, German, Spanish, Russian, and Chinese. Using multilingual encoders keeps architecture and training constant while varying language. The framework and all scripts are released for reproducibility and extension to new languages and models.

Sentence-level lemma substitution (+L). "+L" replaces every token with its UD `LEMMA` (from the CoNLL-U treebank, with identity fallback) and evaluates the lemma of the masked target. Only gold lemmas are used; no external lemmatizer is applied.

Condition	Input string (EN)	Gold
Orig	the[1] scientist[2] [MASK] the[3] books[4] yesterday[5].	analyzed
Orig+L	the[1] scientist[2] [MASK] the[3] book[4] yesterday[5].	analyze
Full	yesterday[5] the[3] the[1] scientist[2] books[4] [MASK].	analyzed
Full+L	yesterday[5] the[3] the[1] scientist[2] book[4] [MASK].	analyze
Part	the[1] books[4] scientist[2] the[3] yesterday[5] [MASK].	analyzed
Part+L	the[1] book[4] scientist[2] the[3] yesterday[5] [MASK].	analyze
Head	the[1] scientist[2] books[4] the[3] [MASK] yesterday[5].	analyzed

Table 1: Illustrative English examples for each condition. Bracketed indices [i] mark token identity for expository purposes only. "+L" applies sentence-level UD lemma substitution; the gold label is the lemma of the masked target.

Word-order perturbations. Three complementary perturbations probe different aspects of word-order sensitivity: (i) maximal disruption of sequence information, (ii) disruption that keeps local function-word scaffolding, and (iii) dependency-aware disruption targeting predicate–argument anchors. All permutations are deterministic given a (seed, sentence ID) pair; within each seed, the same perturbation is reused across models and conditions.

- **Full**: randomly permutes all UD word tokens, serving as a stress test of positional reliance.
- **Part**: permutes content words (NOUN, PROPN, VERB, ADJ, ADV) while keeping function words fixed, testing whether local scaffolding can partially preserve predictions.
- **Head**: iterate heads in left-to-right surface order; for each head (skipping heads with UPOS=PUNCT), select one eligible dependent (excluding UPOS=PUNCT), using a deterministic RNG keyed by (seed, sentence ID, head index), and swap the surface positions of the head and the selected dependent in the token sequence. Swaps are applied sequentially to a working copy of the token list; the UD tree is not updated.

Function words are identified by Universal POS (UPOS) tags as the complement of {NOUN, PROPN, VERB, ADJ, ADV}. Punctuation (UPOS = PUNCT) is never permuted. A small language-specific stoplist (for example, articles or auxiliaries) serves only as a safety filter and does not override UPOS definitions.

Evaluated conditions. We evaluate Orig, Full, Part, and Head, plus +L counterparts for Orig, Full, and Part. Definitions and examples appear in Table 1.

Masking protocol. Masking is applied before perturbation. For each sentence, one content word is selected and all of its subword pieces are replaced by the model's mask token ([MASK] for mBERT, <mask> for XLM-R). This guarantees that the same target token is evaluated across conditions; under scrambling the mask moves with the permutation. Scoring is done at the word level by reconstructing the full span from subword predictions. Punctuation (UPOS = PUNCT) and closed-class stoplist items are never selected as targets.

Randomization and runs. We run three seeds (1, 2, 3) and average metrics across seeds. For each seed, we deterministically seed (i) target selection and (ii) each perturbation using a function of (seed, sentence ID), so that within a seed the same masked targets and permutations are reused across models and conditions (preventing sampling drift in cross-condition/model comparisons). Across seeds, targets and/or permutations can differ, providing a robustness check against sampling variance. Python, NumPy, and PyTorch are seeded on CPU and GPU for each run.

Scoring. mBERT uses WordPiece tokenization and XLM-R uses SentencePiece/BPE. Predicted words are reconstructed from subword outputs using model-specific detokenization rules. A prediction is correct only if all subword pieces match the gold target exactly after Unicode NFKC normalization, case folding for alphabetic scripts, and punctuation removal.

Reconstruction span cap. We cap word reconstruction at six subword pieces (max_span_pieces = 6) to prevent degenerate targets. This has negligible impact on English, German, Spanish, Russian, or Chinese (balanced N per condition ≈ 300).

Ranking-based evaluation (top-k). Exact word reconstruction (top-1) can understate partial knowledge when perturbations increase uncertainty without fully removing evidence. Therefore, in addition to word-level top-1 accuracy, we report word-level top-5 accuracy (word_at_5) from the same stored top-k candidate lists in our released JSONL outputs: an item is counted correct at top-5 if the gold word appears anywhere in the model's top-5 reconstructed candidates for the masked span. For multi-piece targets, the top-5 list is defined over complete reconstructed *words* (full-span reconstructions produced by the same detokenization routine as top-1), not over individual subword pieces. This complements top-1 by capturing cases where the gold target remains highly ranked even when it is not the single most probable prediction.

Data sampling. We use Universal Dependencies (UD) treebanks (Nivre et al., 2016) for all five languages. Sampling is controlled by fixed seeds,

and the same sentence IDs are used across models. We report both unbalanced sets (all valid items) and balanced sets (downsampled to the smallest per-condition count per language) to enable like-for-like comparisons.

Released artifacts. The release includes perturbation scripts, per-language sentence ID lists, balanced and unbalanced sampling metadata, configuration files, and raw JSONL outputs for each run. All code and metadata are licensed for reuse; UD text is not redistributed.

4. Datasets and Sampling

We evaluate on five UD treebanks spanning diverse typological profiles: English (EWT), German (GSD), Spanish (AnCora), Russian (SynTagRus), and Chinese (GSD). Typologically, English and Chinese are predominantly fixed SVO with minimal/isolating morphology; Spanish and German are fusional with richer inflection (German has case-marking and V2/SOV-in-subordinates); Russian is highly case-marked with relatively free word order. We do not perform any additional training or fine-tuning; all results are inference-time perturbation tests on UD sentences.

Sentence selection. We sample up to 400 sentences per language to bound compute and keep per-language uncertainty comparable (preventing larger treebanks from dominating). We keep the same sentence IDs across models and seeds. For each seed, a single content word is selected and masked per sentence (Sec. 3); each sentence then receives every perturbation condition (Original, Original+L, and scrambling variants with/without +L).

For Turkish, the final balanced test set is smaller than 250 sentences because many candidates violate piece-length constraints for word-level masking and strict reconstruction; together with fewer sentences that meet all perturbation constraints across conditions, this reduces the overlap needed for balanced downsampling.

Balanced vs. unbalanced reporting. Because some conditions can be dropped by filtering (e.g., target span too long, no movement under partial scramble), the raw per-condition counts may differ. We therefore report two views: (i) *unbalanced* (all valid items retained), and (ii) *balanced* (per language, we downsample each condition to the smallest per-condition count). Unless otherwise noted, aggregate figures are from the balanced view to ensure like-for-like comparisons across conditions.

Lang	+L tok. chg	Full pos. chg	Part pos. chg	Head pos. chg
DE	0.344	0.910	0.173	0.339
EN	0.191	0.906	0.160	0.455
ES	0.350	0.945	0.195	0.452
RU	0.555	0.935	0.268	0.570
ZH	0.007	0.950	0.272	0.478

Table 2: Perturbation magnitude on evaluated UD sentences (non-punctuation tokens). +L tok. chg: fraction of tokens with FORM ≠ LEMMA. Pos. chg: fraction of tokens whose absolute position differs from Orig.

Lang	N (mBERT)	N (XLM-R)
DE	300	300
EN	290	304
ES	300	304
RU	286	304
ZH	300	304

Table 3: Balanced per-condition counts (N) for five languages.

Perturbation magnitude (surface change). To contextualize the relative strength of each manipulation, we quantify how much each perturbation changes the surface token sequence. For +L we report the fraction of non-punctuation tokens whose surface form differs from its lemma (token change rate). For scrambling conditions we report the fraction of non-punctuation tokens that change absolute position relative to the original sentence (position change rate). These magnitude differences emphasize that +L and scrambling are not matched manipulations; our conclusions, therefore, concern model sensitivity under these specific perturbations rather than equalized information removal.

Realized balanced sample sizes. Balanced item counts *per condition* for each language-model pair are shown in Table 3. For each language and model, we downsample all conditions to the smallest valid per-condition count in that language (one masked target per sentence), so N can differ across models because validity is model-dependent: the six-piece span cap and reconstruction filters operate on model-specific subword segmentations (WordPiece vs. SentencePiece/BPE), so the intersection of items that survive *all* conditions can be smaller for one model than the other (e.g., a target that is 7 WordPiece pieces but 5 SentencePiece pieces). Variation across seeds was negligible. Turkish yields markedly smaller N because multi-piece targets and strict word-level reconstruction make many items ineligible across conditions, shrinking the intersection needed for balancing.

Tokenization and counting. All counting is sentence-based (one masked word per sentence). We report accuracy at the word level by masking and predicting *entire* word spans (combining all subword pieces; see Sec. 3), ensuring comparability between WordPiece (mBERT) and SentencePiece/BPE (XLM-R).

4.1. Reproducibility and Code Availability

Sampling and permutations are controlled by fixed seeds. For a given seed, the same sentence IDs, masked targets, and perturbation permutations are used across models and conditions. We release the code, sentence lists, condition assignments, and per-run JSONL outputs with the paper.[1] Target selection does not pre-filter by span length; the cap is enforced at scoring time. The punctuation sets and closed-class stoplists used as safety filters are released with the code.

4.2. Pretrained Models and Tokenization

mBERT The multilingual BERT (Devlin et al., 2019) is a transformer-based MLM pretrained on Wikipedia corpora from 104 languages. It has 12 transformer layers, 768 hidden units, and 12 attention heads (110M parameters total). Tokenization is performed with a shared WordPiece vocabulary of 119,547 subword units across languages. The MLM objective masks 15% of input tokens, with the model trained to recover the original tokens from context. We use the original, unfine-tuned model. Note: pretraining coverage is uneven across languages (Wikipedia size varies), so cross-language baselines can reflect both modeling and data exposure; we therefore interpret such differences cautiously (see Sec. 9).

XLM-R XLM-RoBERTa base (Conneau et al., 2020) is a transformer-based MLM pretrained on CommonCrawl-based CC-100 corpora from 100 languages, with substantially larger pretraining data than mBERT. The base architecture also has 12 layers, 768 hidden units, and 12 attention heads (270M parameters total), but uses a shared SentencePiece/BPE vocabulary of 250,002 subword units. Like mBERT, it is trained with a 15% masking rate, but uses the `<mask>` token rather than `[MASK]`. As with mBERT, CC-100 coverage is imbalanced across languages; tokenization and pretraining exposure can therefore contribute to cross-language differences (see Sec. 9).

Masking and scoring. Masking is performed at the *word level* to ensure comparability between WordPiece (mBERT) and SentencePiece/BPE (XLM-R). The target word is identified in the untokenized sentence, and all of its subword pieces are replaced by the model's mask token (`[MASK]` or `<mask>`). Perturbations are applied to the masked input, and each sentence contains exactly one masked content word. Predictions are reconstructed to full words from subword outputs with a cap of *six* pieces; items exceeding the cap are excluded from word-level evaluation. A prediction is counted correct only if the reconstructed form exactly matches the gold target *after Unicode NFKC normalization, case-folding for alphabetic scripts, and punctuation stripping*. The gold target is the *surface form* for non-`+L` conditions and the *lemma* for `+L`.

5. Baseline

Our baseline is the original, unfine-tuned pretrained model evaluated on unmodified sentences (Orig condition) with a single masked target word. For each sentence in the test set, we select one content word from the untokenized text and replace *all* of its subword pieces with the model's mask token (`[MASK]` for mBERT, `<mask>` for XLM-R). This word-level masking is deterministic within a seed: the same *target word* is used across models and conditions; under scrambling its position changes with the permutation. Across seeds, target selection may differ.

Predictions are generated with the model's masked language modeling head. For multi-piece targets, all predicted subwords must match in order for the reconstruction to be marked correct. Accuracy is computed at the *word level* (exact match of reconstructed form after Unicode normalization and case-folding for alphabetic scripts). This fixed-mask baseline serves as the reference point for measuring degradation under the perturbation conditions described in Sec. 3.

6. Results

We report balanced word-level accuracy by language and condition for mBERT and XLM-R. Main-text tables include DE/EN/ES/RU/ZH; Turkish (TR) is reported in Appendix A due to floor effects under strict reconstruction. Conditions: Orig (original), Full (full permutation), Part (subset of content words permuted), Head (each UD head is swapped with one of its dependents in a single pass (seed-deterministic); other tokens remain in place); +L denotes *sentence-level* UD lemma substitution. Values are averaged over seeds on the balanced sets. We report 95% confidence intervals: Wilson CIs for accuracy and parametric boot-

[1] https://github.com/bondfeld/WordPrediction_LREC2026

mBERT							
Lang	Orig	Full	Part	Head	Orig+L	Full+L	Part+L
DE	0.143	0.013	0.075	0.078	0.060	0.003	0.033
EN	0.219	0.016	0.123	0.060	0.167	0.019	0.090
ES	0.251	0.013	0.143	0.063	0.143	0.003	0.085
RU	0.276	0.008	0.119	0.063	0.119	0.008	0.061
ZH	0.443	0.023	0.223	0.133	0.448	0.023	0.230

Table 4: Balanced word-level accuracy for mBERT. "+L" = sentence-level UD lemma substitution (context and label).

XLM-R							
Lang	Orig	Full	Part	Head	Orig+L	Full+L	Part+L
DE	0.338	0.013	0.195	0.168	0.150	0.008	0.078
EN	0.303	0.006	0.166	0.079	0.218	0.003	0.123
ES	0.338	0.013	0.175	0.100	0.181	0.010	0.095
RU	0.293	0.013	0.147	0.078	0.126	0.015	0.068
ZH	0.325	0.025	0.150	0.110	0.333	0.025	0.150

Table 5: Balanced word-level accuracy for XLM-R. "+L" = sentence-level UD lemma substitution (context and label).

strap (2,000 draws) for sensitivities S and interaction I (see Figs. 1, 2, 3).

In addition to top-1 exact reconstruction, we report top-5 word accuracy to capture whether the gold target remains among high-ranked candidates under perturbation (Appendix B). Top-5 results preserve the main conclusion: full scrambling collapses performance even when allowing multiple candidates.

To quantify perturbation impact, we use sensitivity to measure the relative loss from *Orig*:

$$S_{\text{cond}} = \frac{A_{\text{Orig}} - A_{\text{cond}}}{A_{\text{Orig}}}$$

and interaction to measure deviation from additivity for *cond*+L:

$$I_{\text{cond}} = A_{\text{cond+L}} - \left(A_{\text{cond}} + A_{\text{Orig+L}} - A_{\text{Orig}}\right).$$

Note that I is computed in accuracy space and depends on the label definition. Because I mixes *surface-label* terms (Orig, Full) with *lemma-label* terms (Orig+L, Full+L), it is sensitive to that choice. We report I on the balanced sets in Tables 6-7 and include unbalanced counterparts in the appendix.

Interpreting I (overlap vs. synergy). Let $D_X = A_{\text{Orig}} - A_X$ denote harm (in particular $D_{\text{+L}} = A_{\text{Orig}} - A_{\text{Orig+L}}$). Then $I_{\text{full}} = D_{\text{Full}} + D_{\text{+L}} - D_{\text{Full+L}}$. That is, the interaction measure computes the difference between combining the two types of manipulations to using each of them individually. Thus $I > 0$ indicates *sub-additive* harm in accuracy space; outside floor/ceiling regimes this pattern is compatible with overlapping/partially redundant cues,

mBERT					
Lang	S_{full}	S_{part}	S_{head}	$S_{\text{+L}}$	I_{full}
DE	0.912	0.474	0.456	0.579	0.073
EN	0.925	0.438	0.725	0.238	0.055
ES	0.950	0.430	0.750	0.430	0.098
RU	0.972	0.569	0.771	0.569	0.157
ZH	0.949	0.497	0.701	-0.011	-0.005

Table 6: Balanced sensitivities and interaction for mBERT computed in *accuracy* space.

XLM-R					
Lang	S_{full}	S_{part}	S_{head}	$S_{\text{+L}}$	I_{full}
DE	0.963	0.422	0.504	0.556	0.183
EN	0.982	0.450	0.739	0.279	0.082
ES	0.963	0.481	0.704	0.467	0.155
RU	0.957	0.500	0.733	0.569	0.169
ZH	0.923	0.538	0.662	-0.023	-0.008

Table 7: Balanced sensitivities and interaction for XLM-R in *accuracy* space.

whereas $I < 0$ indicates *supra-additive* harm (synergy/complementarity).

Why these metrics. Relative-drop sensitivities S factor out language-model baseline differences A_{Orig}, enabling comparisons across typologically diverse settings. We compute the interaction I in *accuracy* space to test (non-)additivity: $I < 0$ means the combined perturbation (Full+L) harms *more* than the additive expectation (supra-additive harm); $I \approx 0$ is additive; $I > 0$ indicates *sub-additive* effects in accuracy space. Because both S and I depend on A_{Orig} and are bounded by floor/ceiling effects, floor cases (e.g., Turkish) compress dynamic range; accordingly, we also report unbalanced results in the appendix and avoid over-interpreting I when baselines are near zero.

7. Findings and Discussion

Because our scrambling manipulations inevitably conflate loss of linear order with reduced well-formedness, observed effects reflect both factors. We report balanced counts and note floor cases in the appendix; isolating "well-formed-only" perturbations is left for future work.

Word order dominates. Across balanced sets, *full* scrambling drives word-level accuracy close to zero for both models in every language (Tables 4-5; Fig. 1). *Partial* and *head* scrambling also yield

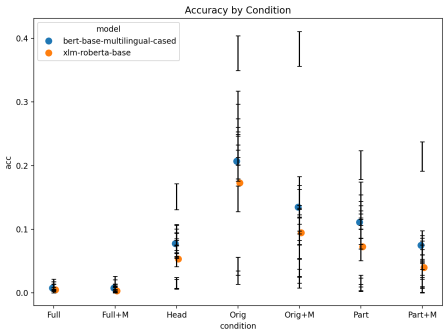

Figure 1: Accuracy across perturbation conditions with Wilson 95% CIs.

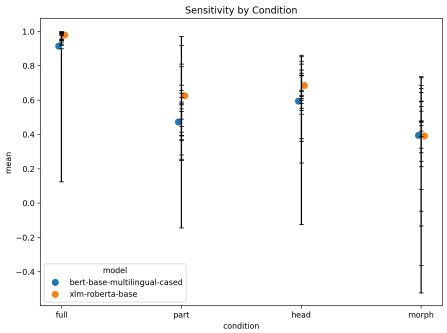

Figure 2: Relative-drop sensitivities S (95% CIs).

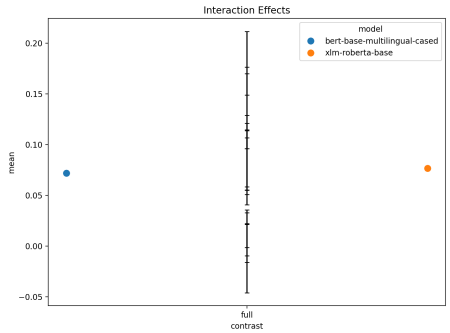

Figure 3: Interaction effects I_{full} (95% CIs).

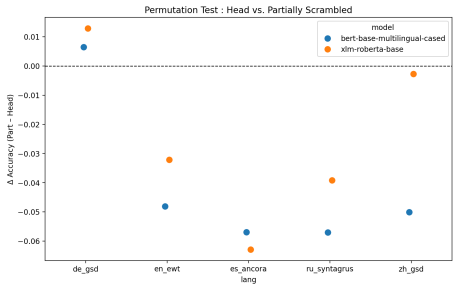

Figure 4: Head vs. Part (Δ accuracy = Part – Head).

large losses (Fig. 2). This conclusion holds under top-5 evaluation as well: even allowing five candidates, Full scrambling yields very low accuracy across languages (at most 0.053 for mBERT and 0.067 for XLM-R on the balanced sets), indicating that perturbations often push the gold target out of the top candidate set rather than only lowering its rank.

English and Chinese show especially large losses under structured disruption (notably *Head*); German is comparatively less affected in several structured conditions, while Russian remains highly order-sensitive under *Full* despite rich case marking. Within the structured perturbations, *Head* tends to hurt English more than *Part*, while German is comparatively less affected, consistent with available morphological and function-word cues. We interpret claims conservatively in floor regimes (e.g., strict span reconstruction in Turkish).

Scramble + Lemma (+L) is sub-additive in accuracy space (with a floor caveat). In DE/EN/ES/RU, I_{full} is positive (Tables 6-7), i.e., accuracy under *Full+L* is higher than the additive baseline $A_{\text{Full}} + A_{\text{Orig+L}} - A_{\text{Orig}}$. However, because *Full* drives accuracy close to zero, the additive baseline is often negative, and I_{full} is mechanically biased upward by the 0-floor; we therefore interpret I_{full} cautiously as a coarse non-additivity diagnostic rather than a clean measure of cue overlap. Consistent with this, on the unbalanced sets the corresponding *Part* interactions are also positive in DE/EN/ES/RU (Appendix Table 10), where floor effects are less severe. In ZH, +L is nearly identity (Table 2), and I_{full} is near zero.

Heads vs. partial content shuffles. Head–dependent disruption affects languages asymmetrically. English shows a larger Head vs. Part gap than German (consistent with fixed SVO and light inflection). In most other languages (ES/RU/ZH), *Head* is also more harmful than *Part* for both models (Fig. 4). German exhibits a small exception for mBERT (Head $\approx$ Part), while for XLM-R, *Head* is slightly more harmful than *Part*.

Why can I_{full} be positive? Under *Full* scrambling, A_{Full} is often near 0, so the additive baseline $A_{\text{Full}} + A_{\text{Orig+L}} - A_{\text{Orig}}$ can fall below 0; because accuracy is bounded below by 0, I_{full} is mechanically biased upward in these floor regimes. We therefore treat I_{full} as a coarse non-additivity diagnostic and corroborate it with less floor-bound interactions (e.g., I_{part} on unbalanced sets; Appendix Table 10). Outside strong floor/ceiling regimes, positive I is compatible with partially overlapping evidence from word order and inflectional form, but we avoid reading $I_{\text{full}} > 0$ as a clean cue decomposition in the near-collapse setting.

Why is $I_{full} > 0$?

Why does +L help (slightly) in ZH for mBERT? In Chinese, lemma substitution is vacuous because

surface forms ≈ lemmas. The tiny uptick (Orig $0.443 \rightarrow 0.448$; Full unchanged at 0.023) reflects tokenization/normalization quirks in mBERT, not genuine morphology.

Lemma substitution (+L, sentence-level). In +L we replace *every token* with its UD LEMMA (CoNLL-U; identity fallback), and evaluate against the lemma of the masked target. Thus, +L lemmatizes both context and target.

Model-specific notes. (1) Chinese. On the balanced sets, both models achieve substantive baselines (mBERT 0.443, XLM-R 0.325; Tables 4-5). +L has negligible effect overall, and I_{full} is near zero or slightly negative (Tables 6-7), suggesting limited supra-additivity between order and lemma substitution under our reconstruction protocol. (2) Turkish. Both models are near floor even on Orig; every sensitivity/interaction statistic on TR should be read with floor effects in mind.

What this means. These multilingual MLMs are still positional workhorses. Morphology on the target helps a bit (sometimes more in DE/ES/RU), but it does *not* compensate for lost order. Head ordering of core arguments matters a lot in English; German's morphology together with function words makes head swaps slightly less harmful. Cross-model differences (e.g., ZH) are large enough that claims about "the model" need to specify which one.

Lessons learned. Two themes stand out. First, non-additivity: in DE/ES/RU, we observe $I_{\text{full}} > 0$ for both models (e.g., XLM-R: $0.183/0.155/0.169$; mBERT: $0.073/0.098/0.157$), meaning the loss under *Full+L* is smaller than the sum of the separate losses from *Full* and *+L*. This is consistent with overlapping/partially redundant signals from word order and the target's morphology rather than complementary amplification. Second, head-dependent disruption is asymmetric across languages: English is hit especially hard (fixed SVO, light inflection), whereas German shows mild cushioning. In ES/RU/ZH, *Head* typically hurts more than *Part* for both models; in German this holds for XLM-R (Part 0.195 vs. Head 0.168) but not for mBERT (Part 0.075 vs. Head 0.078, i.e., roughly equal). In Chinese, unlike the other languages, *+L* does not reduce the baseline (mBERT: $0.443 \rightarrow 0.448$; XLM-R: $0.325 \rightarrow 0.333$); we attribute these tiny upticks to tokenization/reconstruction effects rather than genuine morphological leverage, and *+L* never compensates for the loss of order. Finally, Turkish is dominated by floor effects: very low *Orig* baselines and balancing-induced reductions in N compress dynamic range, rendering S and I numerically unstable; we therefore defer to unbalanced summaries in Appendix A.

8. Conclusion

Across five languages and two multilingual MLMs (mBERT, XLM-R), word order is the dominant cue. Full scrambling causes near-total collapse in masked word prediction (relative drops ≈ 0.91-0.98 across languages/models); partial and head-only scrambling also inflict large losses. Keeping function words in place does not rescue predictions, and shuffling the relative order of core heads is particularly harmful in English, while German is slightly cushioned by function-word/morphological scaffolding.

+L removes information and generally reduces accuracy; it never compensates for lost order. Any apparent improvements (e.g., tiny upticks for mBERT-ZH) are evaluation artifacts from tokenization/reconstruction, not genuine gains. Interactions computed in accuracy space are mostly positive in DE/ES/RU ($I_{\text{full}} > 0$), indicating sub-additivity (joint harm $<$ additive baseline). Because *Full* accuracy is often near floor, I_{full} is partly shaped by saturation; we therefore treat it as a coarse non-additivity signal and corroborate it with less floor-bound interactions (e.g., I_{part} on unbalanced sets; Appendix Table 10).

Baseline behavior matters. mBERT and XLM-R diverge sharply on some languages (notably Chinese, where mBERT's baseline is much higher), and Turkish baselines are so low that "drops" and interaction scores are floor-limited. The evidence indicates strong behavioral sensitivity to word-order disruption in this diagnostic setting and only limited robustness under lemma-normalized contexts.

Two directions look promising: (i) training objectives that reduce over-reliance on absolute order (e.g., de-scrambling or order-invariant auxiliaries) while preserving syntactic signals, and (ii) morphology-aware prediction heads that expose inflectional structure explicitly rather than relying on subword artifacts. Evaluation-wise, adding constituency-based scrambling and human baselines on the same items would sharpen what cues models vs. humans actually use when order is unreliable.

9. Limitations

We evaluate *Orig*, *Full*, *Part*, and *Head*, plus *+L* variants where used (*Orig+L*, *Full+L*, *Part+L*); *+L* is *sentence-level* lemma substitution: the whole context is lemmatized and the gold label is the lemma of the masked target. We do not include target-only or context-only normalization, and we do not evaluate Head+L.

Exactly one content word is masked per sentence. Within each seed, the target is fixed across

conditions (and across models), but targets can differ across seeds. Accuracy is exact word-level reconstruction from subword predictions, which penalizes near-misses and interacts with subword length/frequency, potentially understating partial knowledge. Tokenization differs by model (WordPiece for mBERT; SentencePiece/BPE for XLM-R) and across languages, so some baseline gaps and asymmetries likely reflect segmentation as well as modeling. We partially address this by reporting top-5 word accuracy; however, we do not report full probability shifts (e.g., NLL/entropy) because we do not store full vocab distributions for each item. Head-only scrambling uses UD *gold* trees: in one seed-deterministic pass we swap each head token with one of its dependents (any label), leaving other tokens fixed. We do not parse at test time and do not perform phrase/constituency-level reordering.

For each language-model pair we downsample every condition to the smallest per-condition count, which reduces N; Turkish is most affected because many targets exceed the six-piece reconstruction cap, producing floor effects that make sensitivities and interactions numerically fragile. We report means on balanced sets and compute sensitivities/interactions in *accuracy* space; given the near-collapse under *Full* scrambling, additional significance tests add little beyond visible effect sizes, so we omit p-values.

Finally, this is a diagnostic self-contained masked-word recovery study with unfine-tuned models on UD sentences; it is distribution-shifted relative to pretraining and may not predict downstream, fine-tuned behavior. We evaluate two base multilingual encoders (mBERT, XLM-R); extending to newer multilingual LMs is left for future work. Our protocol (balanced sampling, structured scrambling, accuracy-space interactions) transfers directly.

Acknowledgments

This research was supported in part by the National Science Foundation under Grant No. *2226006*.

10. Bibliographical References

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019. Cross-lingual dependency parsing with unlabeled auxiliary languages. *CoRR*, abs/1909.09265.

Xuanda Chen, Timothy O'Donnell, and Siva Reddy. 2024. When does word order matter and when doesn't it?

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Alexander Fung, Chengxu Zhuang, Steven T. Piantadosi, Jacob Andreas, and Evelina Fedorenko. 2024. Word-order error detection helps data-efficient language models learn syntax.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Prakhar Gupta and Martin Jaggi. 2021. Obtaining better static word embeddings using contextual embedding models. 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 5241–5253, Online. Association for Computational Linguistics.

Jack Hessel and Alexandra Schofield. 2021. How effective is BERT without word ordering? implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 521–535.

Zoey Liu. 2020. Mixed evidence for crosslinguistic dependency length minimization. *STUF - Language Typology and Universals*, 73(4):605–633.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. *Language Resources and Evaluation*, 50(1):165–210.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? 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 851–864, Online. Association for Computational Linguistics.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, BERT doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, Dublin, Ireland. Association for Computational Linguistics.

Thuong-Hai Pham, Dominik Machácek, and Ondrej Bojar. 2019. Promoting the knowledge of source syntax in transformer NMT is not needed. *CoRR*, abs/1910.11218.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pretraining for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Benyou Wang, Donghao Zhao, Christina Lioma, Qiuchi Li, Peng Zhang, and Jakob Grue Simonsen. 2019. Encoding word order in complex embeddings. *CoRR*, abs/1912.12333.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019. Convolutional self-attention networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4040–4045, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Appendix A: Turkish (TR) Results

Scope. This appendix reports (i) per-language balanced accuracies and N's, (ii) unbalanced summaries to avoid floor/ceiling compression, and (iii) Turkish (TR) results, which are floor-limited under strict word-level reconstruction.

Conventions. Word-level accuracy is exact match after Unicode NFKC and case-folding (alphabetic scripts). Confidence intervals are Wilson 95% CIs for accuracy and parametric 95% CIs for sensitivities/interactions (2,000 bootstrap draws). "+L" denotes *sentence-level* UD lemma substitution (all tokens lemmatized; target label is the lemma; identity fallback when missing).

Filtering and balancing. Each sentence contains one masked content word; items exceeding the six-piece reconstruction cap are excluded. For "Part," examples with no movement under the content-word shuffle are dropped. Balanced sets downsample each condition to the smallest per-condition count within a language–model pair; unbalanced tables include all valid items.

Computation notes. Sensitivities are $S_{\text{cond}} = (A_{\text{Orig}} - A_{\text{cond}})/A_{\text{Orig}}$. Interaction is $I_{\text{cond}} = A_{\text{cond+L}} - (A_{\text{cond}} + A_{\text{Orig+L}} - A_{\text{Orig}})$, computed in accuracy space. Head scrambling swaps each UD head token with one dependent in a single, seed-deterministic pass; other tokens remain fixed.

Background. Turkish (TR) exhibits strong floor effects because agglutinative morphology yields long, multi-piece targets that often exceed the six-piece reconstruction cap. After filtering, the remaining items sit near floor. We emphasize unbalanced TR summaries and interpret absolute differences cautiously.

	N (mBERT)	N (XLM-R)
TR (balanced per condition)	108	152

Table 8: Balanced per-condition counts (N) for Turkish (TR).

TR: Balanced word-level accuracy							
Model	Orig	Full	Part	Head	Orig+L	Full+L	Part+L
mBERT	0.000	0.000	0.019	0.006	0.013	0.013	0.019
XLM-R	0.063	0.000	0.057	0.013	0.006	0.006	0.019

Table 9: Balanced word-level accuracies for Turkish (TR). Interpret with caution (floor, selection, reconstruction effects).

Note on Table 9. In TR we sometimes see $A_{\text{cond}} > A_{\text{Orig}}$; with near-floor baselines and balanced downsampling (span-cap and "no-move" filters), these tiny deltas are noise rather than real gains from scrambling.

Table 10: Sensitivities and interaction in *accuracy* space on *unbalanced* sets (all languages shown for cross-reference).

mBERT						
Lang	S_{full}	S_{part}	S_{head}	S_{morph}	I_{full}	I_{part}
DE	0.980	0.418	0.438	0.559	0.081	0.036
EN	0.954	0.457	0.706	0.323	0.073	0.026
ES	0.964	0.420	0.712	0.407	0.094	0.042
RU	0.982	0.533	0.724	0.629	0.166	0.096
TR	0.719	0.450	0.465	0.143	0.004	0.000
ZH	0.958	0.490	0.655	-0.005	0.000	0.005
XLM-R						
Lang	S_{full}	S_{part}	S_{head}	S_{morph}	I_{full}	I_{part}
DE	0.968	0.470	0.563	0.594	0.173	0.084
EN	0.964	0.445	0.716	0.297	0.079	0.033
ES	0.977	0.498	0.715	0.458	0.148	0.072
RU	0.975	0.546	0.723	0.604	0.168	0.096
TR	0.931	0.437	0.706	0.688	0.045	0.017
ZH	0.938	0.587	0.644	0.012	0.004	0.001

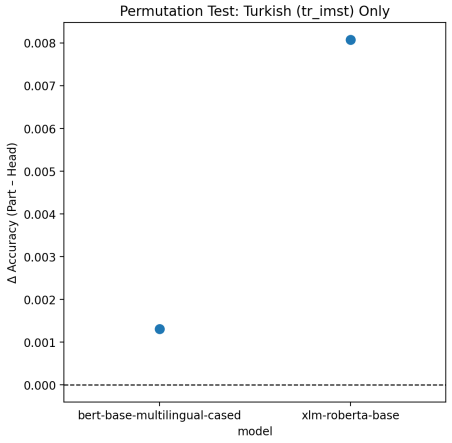

Figure 5: Turkish (TR) only: Head vs. Part (Δ accuracy = Part – Head).

Interpretation. Both models are at or near floor in TR across all conditions. Because Orig accuracy is close to zero, relative drops and interaction terms are not meaningful indicators of robustness. Micro-differences between conditions are within noise and largely reflect span-cap filtering rather than genuine morphological or order sensitivity.

Appendix B

mBERT (Top-5)							
Lang	Orig	Full	Part	Head	Orig+L	Full+L	Part+L
DE	0.282	0.021	0.185	0.153	0.166	0.013	0.102
EN	0.400	0.031	0.236	0.140	0.312	0.029	0.191
ES	0.423	0.033	0.264	0.170	0.266	0.037	0.175
RU	0.403	0.014	0.201	0.127	0.172	0.017	0.105
ZH	0.588	0.053	0.331	0.247	0.583	0.055	0.328

Table 11: Balanced top-5 word accuracy for mBERT.

XLM-R (Top-5)							
Lang	Orig	Full	Part	Head	Orig+L	Full+L	Part+L
DE	0.503	0.030	0.311	0.289	0.281	0.018	0.171
EN	0.485	0.031	0.288	0.184	0.390	0.022	0.252
ES	0.496	0.030	0.294	0.233	0.300	0.022	0.193
RU	0.439	0.022	0.262	0.184	0.224	0.025	0.153
ZH	0.464	0.067	0.263	0.237	0.466	0.066	0.263

Table 12: Balanced top-5 word accuracy for XLM-R.